Behav Res (2015) 47:1274–1294
DOI 10.3758/s13428-014-0527-2

# Meta-analysis of multiple outcomes: a multilevel approach

**Wim Van den Noortgate · José Antonio López-López · Fulgencio Marín-Martínez · Julio Sánchez-Meca**

Published online: 1 November 2014


**Abstract** In meta-analysis, dependent effect sizes are very common. An example is where in one or more studies the effect of an intervention is evaluated on multiple outcome variables for the same sample of participants. In this paper, we evaluate a three-level meta-analytic model to account for this kind of dependence, extending the simulation results of Van den Noortgate, López-López, Marín-Martínez, and Sánchez-Meca *Behavior Research Methods*, 45, 576–594 (2013) by allowing for a variation in the number of effect sizes per study, in the between-study variance, in the correlations between pairs of outcomes, and in the sample size of the studies. At the same time, we explore the performance of the approach if the outcomes used in a study can be regarded as a random sample from a population of outcomes. We conclude that although this approach is relatively simple and does not require prior estimates of the sampling covariances between effect sizes, it gives appropriate mean effect size estimates, standard error estimates, and confidence interval coverage proportions in a variety of realistic situations.

**Keywords** Multilevel · Meta-analysis · Multiple outcomes · Dependence

A complicating factor in meta-analysis is dependence in the observed measures of association. Dependence is found in a wide variety of applications and research areas, such as described and illustrated by Jackson, Riley, and White (2011). A common source of dependence is the occurrence within studies of multiple observed measures of association that are calculated using the same study sample. A study might for instance explore the effect of a therapy on two different constructs, such as anxiety and depression, or might use two different instruments to measure a specific construct (e.g., two depression inventories). If two outcome variables are correlated, then so are the measures of association between the outcomes and an independent variable. For instance, suppose that the difference between a control group and an experimental group is evaluated for two different but related outcomes (outcome $j$ and $j'$ respectively) by means of standardized mean differences (SMDs; $d_j$ and $d_{j'}$), each equal to the difference in sample means for both conditions, divided by the square root of the pooled within group variance, $s_p$. Gleser and Olkin (1994) stated that the sampling variance of $d_j$ and the sampling covariance between $d_j$ and $d_{j'}$, are equal to

$$\sigma^2_{d_j} \cong \frac{n_E + n_C}{n_E n_C} + \frac{\delta_j^2}{2(n_E + n_C)} \tag{1}$$

and

$$\sigma_{d_j d_{j'}} \cong \frac{n_E + n_C}{n_E n_C}\rho_{jj'} + \frac{\delta_j \delta_{j'} \rho^2_{jj'}}{2(n_E + n_C)}, \tag{2}$$

with $n_E$ and $n_C$ referring to the size of the experimental and control groups, respectively, $\delta_j$ and $\delta_{j'}$ to the population SMDs estimated by $d_j$ and $d_{j'}$, and $\rho_{jj'}$ to the correlation between the two outcome variables.

Equation 2 shows that if two outcome variables are positively correlated, the sampling covariance between the corresponding SMDs is also positive. This means that if in a specific sample for one outcome the observed SMD is larger

W. Van den Noortgate (✉)
Faculty of Psychology and Educational Sciences and itec-iMinds, University of Leuven, Vesaliusstraat 2, 3000 Leuven, Belgium
e-mail: wim.vandennoortgate@kuleuven-kortrijk.be

J. A. López-López
School of Social and Community Medicine, University of Bristol, Bristol, UK

F. Marín-Martínez · J. Sánchez-Meca
Faculty of Psychology, Universidad de Murcia, Murcia, Spain

than the population SMD, we can expect that the observed SMD for the other outcome will also be larger than the population SMD. It is important that this dependence is taken into account when performing a meta-analysis, because ignoring the dependence can result in biased statistical inferences (Becker, 2000). More specifically, by treating effect sizes as independent, we ignore the overlap in information that is delivered by each of the observed effect sizes, resulting in underestimated standard errors. As a consequence, confidence intervals will be too small and the Type I error rate will be inflated when estimating and testing the population effect size. Becker (2000) therefore concludes that "No reviewer should ever ignore dependence among study outcomes. Even the most simplest ad hoc options are better than pretending such dependence does not exist." In a review of 56 meta-analyses in the domain of education, Ahn, Ames, and Myers (2012) found an average of 3.6 effects per studies. Only one meta-analysis explicitly mentioned that dependence was not a problem. In 15 meta-analyses the potential issue of dependence was not addressed. In the other meta-analyses, dependence was dealt with in various ways.

One way to deal with the dependence is performing separate meta-analyses for each type of outcome. A drawback of this approach, however, is that in this way each meta-analysis is done on only a part of the available data, making this approach less efficient. For some outcomes the number of studies might even be too small to yield meaningful results. An often used alternative approach (which was used in 21 of the 56 meta-analyses studied by Ahn et al., 2012) is to aggregate the effect sizes within each study before combining the results over studies. S.F. Cheung and Chan (2014) describe how the effective sample size of the mean effect within each study can be determined, accounting for the intercorrelation between the effect sizes within the study. By aggregating effect sizes, however, we lose information about differences in the effect on different outcomes and as a result it might become impossible to estimate and test moderator effects of outcome characteristics. A combination of both approaches is also possible: a separate meta-analysis is done for each type of outcome, and multiple observed effect sizes for outcomes of the same type are aggregated within studies, before being meta-analyzed over studies. This combined approach was used in about 14 % of the meta-analyses studied by Ahn et al. (2012). Still another approach is to account for the sampling covariance by using multivariate meta-analytic models (see Kalaian & Raudenbush, 1996, and Raudenbush, Becker, & Kalaian 1988, for a description, van Houwelingen, Arends & Stijnen, 2002, and Arends, Voko & Stijnen, 2003, for illustrations and Jackson et al., 2011, for an overview of the state of the art), in which as in a univariate model the precision of the estimates of the population effect sizes is maximized by weighting the observed effect sizes by their estimated precision. However, whereas in a univariate meta-analysis these precision estimates are based on prior estimates of the sampling variance for each observed effect size, in a multivariate meta-analysis not only is the sampling variance of the observed effect sizes taken into account, but also the sampling covariances. An important advantage of the multivariate approach is that in the analysis, data from all outcomes are used at the same time, which means that to estimate the mean effect or the between-study variance for a specific outcome, 'strength is borrowed' from observed effect sizes for other outcomes, therefore resulting in more accurate effect estimates or smaller standard errors (Jackson et al., 2011). Unfortunately, whereas the sampling variance can easily (and quite accurately) be estimated by replacing the population effect size in Equation 1 by the observed effect size, reliable information about the correlations between the outcome variables is only rarely reported in primary studies, and often researcher-designed measures are used to assess the effect of an intervention, making it difficult to obtain a good estimate of the sampling covariance (Scammacca, Roberts & Stuebing, 2013). Whereas Ishak, Platt, Joseph, and Hanley (2008) suggest that if the interest in a multivariate meta-analysis is only on the treatment effects and not on the between-study correlation between outcomes, ignoring the within-study covariation will not distort the results, Riley (2009) showed that this is not generally true and that ignoring the correlations might increase the mean-square error and standard error of the pooled estimates, as well as their bias for non-ignorable missing data. Riley (2009) describes some ways in which this problem could be tackled, for instance by analyzing the raw data from the primary studies, by using external data to narrow down the possible values for the correlation coefficients or by performing sensitivity analyses over the entire correlation range, but as noted by Scammacca et al. (2013) performing sensitivity analyses can easily become too laborious and time-consuming and even not feasible if more than three outcomes are used. A lack of information about these correlations is, in addition to the relative complexity of the multivariate approach, an important reason why the multivariate meta-analysis is not often used. For instance, the multivariate approach was used in none of the 56 meta-analyses that were studied by Ahn et al. (2012).

Besides the problem of the missing between-outcome correlations, using a multivariate approach is less straightforward if there is no consensus in the research area about the most appropriate outcomes, and therefore studies do not report results for more or less the same outcomes. When there is a lot of variation between studies in the outcome variables, using a multivariate model can easily become infeasible because of the large number of correlations that should be 'known' before doing the meta-analysis. As an example, let us look at the meta-analysis of Geeraert, Van den Noortgate, Grietens and Onghena (2004), combining the results of 39 studies evaluating the effect of early prevention programs for

families with young children at risk for physical child abuse and neglect. The authors found considerable variation in the criteria that were used to evaluate the effect of the prevention programs. This is not surprising, because child abuse and neglect are very difficult to observe directly and there is no single measure that can be regarded as a highly reliable indicator of child abuse. More specifically, the authors found that reported outcomes included direct reports of child abuse or neglect by child protective services, as well as indirect indications of child abuse and neglect such as reports of hospitalization, the frequency of medical emergency service visits, contacts with youth protection services and out-of-home placements. A lot of studies also evaluated the reduction of risk, looking at, for instance, the well-being of the child, parent-child interaction characteristics, and social support. Thirty-four studies out of thirty-nine calculated the effect for more than one outcome variable. The total number of effect sizes was 587. The number of effect sizes per study varied from 1 to 52, with an average of about 15.

Hedges, Tipton and Johnson (2010) proposed the robust variance estimation (RVE) approach. In this approach, the dependence is not explicitly modelled, but instead the standard errors for the overall treatment effect or meta-regression coefficients are adjusted. Although in the Hedges et al. approach a reasonable guess of the between outcome correlation is needed to estimate the between-study variance and to approximate the optimal weights, results are shown to be influenced only slightly by the choice of this value.

An alternative approach that does not require prior covariance estimates is a three-level approach. This approach was used by Geeraert et al. (2004), modeling the sampling variation for each effect size (level one), variation over outcomes within a study (level two), and variation over studies (level three). The basic model consists of three regression equations, one for each level:

$$d_{jk} = \beta_{jk} + r_{jk} \; with \; r_{jk} \sim N\left(0,\ \sigma^2_{r_{jk}}\right) \tag{3}$$

$$\beta_{jk} = \theta_{0k} + v_{jk} \; with \; v_{jk} \sim N\left(0,\ \sigma^2_{v}\right) \tag{4}$$

$$\theta_{0k} = \gamma_{00} + u_{0k} \; with \; u_{0k} \sim N\left(0,\ \sigma^2_{u}\right) \tag{5}$$

The equation at the first level, the sample level (Eq. 3), states that $d_{jk}$, the $j^{th}$ observed effect size ($j$ =1, 2, …, $J$) from study $k$ ($k$ =1, 2, …, $K$), is equal to the corresponding population value $\beta_{jk}$, plus a random deviation, $r_{jk}$. The random residuals are supposed to be normally distributed with zero mean and a variance, $\sigma^2_{r_{jk}}$, that might be study- and outcome-dependent, and strongly depends on the size of the study. As in a typical meta-analysis, this variance is estimated before performing the three-level meta-analysis. For instance, the sampling variance for the SMD can be estimated by replacing the population effect size in Equation 1 by its observed counterpart. The equation at the second level, the level of outcomes (Eq. 4), states that the population effects for the different outcomes within a study can be decomposed in a study mean ($\theta_{0k}$) and random residuals ($v_{jk}$), that are again assumed to be normally distributed. At the third level, the study level (Eq. 5), study mean effects are modeled as randomly varying around an overall mean ($\gamma_{00}$). By substitution, we can write Equations 3 to 5 in one single equation:

$$d_{jk} = \gamma_{00} + u_{0k} + v_{jk} + r_{jk} \tag{6}$$

This three-level model is an extension of the random effects model that is commonly used and promoted in meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2010):

$$d_k = \gamma_0 + u_k + r_k \tag{7}$$

Whereas in the traditional random effects meta-analytic model (Eq. 7), observed effect sizes are regarded to vary due to sampling variance ($\sigma^2_{r_k}$) and systematic between-study variance ($\sigma^2_u$), the three-level model (Eq. 6) includes two kinds of systematic variance: variance between studies ($\sigma^2_u$) and variance between outcomes from the same study ($\sigma^2_v$).

In the same way as the traditional random effects model can be extended to a mixed effects model by including study characteristics that possibly explain (part of) the between-study variance, we can extend the three-level random effects model to a three-level mixed effects model by including characteristics of the outcomes (e.g., we could use two dummy variables to model a possible difference within studies between direct reports of child abuse, indirect indications and risk factors) as predictors at the outcome level (as will be done below, see Eq. 10). Possible variation over studies of the new coefficients at the outcome level can be described using additional equations at the study level. Study characteristics (e.g., the kind of prevention program that is evaluated) can be used as predictors in the study level equations, in an attempt to explain differences between studies.

Parameters that are estimated in a multilevel meta-analysis are the regression coefficients of the highest level equations (c.q., $\gamma_{00}$, interpreted as the mean effect), as well as the variances at the second and higher levels (c.q., $\sigma^2_v$, the variation between outcomes within the same study, and $\sigma^2_u$, the variation over studies). Multilevel model parameters are typically estimated and tested using maximum likelihood estimation procedures. Compared to the full maximum likelihood procedure (ML), the restricted maximum likelihood procedure (REML) decreases the bias in the variance component

estimates, but will only yield valid results for the likelihood ratio test when comparing nested models that have the same fixed part. Therefore, only if ML was used, the likelihood ratio test can be used for testing fixed parameters. For a more elaborated description of the multilevel approach to meta-analysis, we refer to Hox (2002), Raudenbush and Bryk (2002), Van den Noortgate and Onghena (2003) and Konstantopoulos (2011). M. W.-L.Cheung (2013) described how the structural equation modelling framework can be used to perform three-level analyses, which provides a flexible approach for testing model constraints, for constructing likelihood-based confidence intervals for the heterogeneity variances and for handling missing data.

A crucial assumption in a multilevel model is that the residuals at each level are independent from each other, from residuals at other levels and from the regression coefficients (Raudenbush & Bryk, 2002). As a result, the residuals from the sample level equation (Eq. 3) – referring to the deviations of the effect sizes of the multiple observed outcomes from the corresponding population effects – are assumed to be independently distributed, more specifically normally distributed with an outcome- and study-specific variance. Therefore, the sample level equation does not take into account a possible sampling covariation that nevertheless can be expected if multiple effects are estimated for the same sample. However, the use of three-level models exactly aims at accounting for covariation between outcomes: because the residuals at each level are independent from each other, from residuals at other levels and from the regression coefficients, it is easy to see that the variance between studies reflects the covariance between two effect sizes from the same study:

$$\sigma_{d_{jk}d_{j'k}} = \sigma_{\left(\gamma_{00}+u_{0k}+v_{jk}+r_{jk}\right)\left(\gamma_{00}+u_{0k}+v_{j'k}+r_{j'k}\right)} = \sigma_{u_{0k}u_{0k}} = \sigma_u^2 \qquad (8)$$

Proof is given in Appendix A. This equality can be understood as follows: a large positive covariance between outcomes from the same study means that if in a study a relatively large effect is observed also other effects in that study are expected to be relatively large, resulting in a relatively large study mean. Covariation between outcomes therefore results in small differences within studies, but large differences between study means. If on the contrary the effect of outcomes is independent of the effect of other outcomes, this will result in differences between effects within studies, but differences between the study mean effects will be relatively small. Because residuals at each level are regarded as independent, only the third level variance component accounts for correlation between outcomes within the same study.

Van den Noortgate, López-López, Marín-Martínez and Sánchez-Meca (2013) showed using extensive simulation that this variance component at the third level reflects the total dependence between outcomes from the same study (both the sampling covariance and the covariance between population effects): the estimated between-study variance was on average very close to the sum of the covariance between two outcomes in population effects and the sampling covariance as calculated using Equation 2. Therefore, although the three-level meta-analysis does not require prior estimates of the sampling covariance, the sampling covariance is accounted for, resulting in correct statistical inferences: standard errors were found unbiased and confidence interval coverage proportions for the mean effect estimate corresponded to their nominal levels.

However, because only one parameter, the variance at the third level, refers to the covariance between effect sizes from the same study, an underlying assumption of the three-level model is that the covariance is the same for all pairs of outcomes and for all studies. Yet, Equation 2 indicates that the covariance depends on the size of the sample, as well as on the correlation between the outcome variables, which is not necessarily the same for all pairs of outcomes. A restriction of the simulation study of Van den Noortgate et al. (2013) is that simulated studies were of the same size and the covariance between outcomes was the same for all pairs of outcomes. Moreover, data were simulated making some additional strict assumptions that often are not realistic in practice: in all meta-analytic datasets, studies reported on the same number of outcomes and the between-study variance was assumed equal for all types of outcomes. The purpose of this article is therefore to explore the performance of the three-level approach in more complex, but more realistic, situations. Moreover, in this paper we will discuss the meta-analysis of effect sizes, rather than on a multilevel analysis of raw data.

The focus of the paper therefore is on the use of a three-level model to model the dependence within studies that is due to measuring multiple outcomes in the same study. we want to contribute to a better understanding of this approach, as called for by M. W.-L. Cheung (2013) and Scammacca et al. (2013). Other studies (Stevens & Taylor 2009; Konstantopoulos, 2011) proposed three-level models to account for dependence over studies, for instance when studies are nested within research groups or school districts. Stevens and Taylor (2009) proposed methods to account for this dependence between studies, combined with a specific kind of dependence in multiple effect sizes from the same study –a kind that is also not the focus of this paper-, dependence that occurs when mean differences from independent groups are standardized using a common within-group standard deviation estimate. Another kind dependence occurs if effect sizes compare multiple treatment groups with a common control group (see e.g., Scammacca et al., 2013, for a description of possible approaches in this scenario).

In the remainder of the paper, we first illustrate the three-level approach by reanalyzing the Geeraert et al. (2004) meta-

analytic data set. Next, we look at the setup and the results of a simulation study. For all analyses and simulations, we used the REML procedure implemented in SAS Proc Mixed (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Codes used for the example are given in Appendix B.

## Example

The results in Table 1 show that using a three-level random effects model (Eq. 6) for the child abuse and neglect data set Geeraert et al. (2004) described above yields a mean effect estimate equal to 0.23 with a standard error equal to 0.026.[1] Using the RVE approach, the estimate of the overall effect equals 0.26 with a standard error of 0.032.

If we would regard effect sizes as independent, as if each observed effect size came from a separate study, the three-level model reduces to an ordinary random effects model (Eq. 9), with only one systematic variance component: whereas in the three-level analysis we estimate the variance between outcomes from the same study and the variance between studies, in the ordinary random effects model we only estimate the variance between all outcomes (regardless of the study they stem from).

$$d_{jk} = \gamma_{00} + v_{jk} + r_{jk} \tag{9}$$

Table 1 shows that by using this two-level model (Eq. 9), we would obtain a somewhat smaller mean effect estimate of 0.20. As expected (Becker, 2000), ignoring the dependence results in a much smaller standard error for this mean effect.

Geeraert et al. (2004) also investigated whether the effect depends on the kind of outcome. Therefore, they defined two dummy variables referring to the kind of outcome: $X_{1jk}$ is equal to one if outcome $j$ from study $k$ refers to an indicator of child abuse or neglect, zero otherwise; $X_{2jk}$ is equal to one if outcome $j$ from study $k$ refers to a risk factor of child abuse or neglect, zero otherwise. To estimate the mean effect for both types of outcomes, both dummy variables are included as predictors in the model and the intercept is dropped:

$$d_{jk} = \gamma_{100}X_{1jk} + \gamma_{200}X_{2jk} + u_{0k} + v_{jk} + r_{jk} \tag{10}$$

An equivalent model is a model with an intercept but with only the first (or second) dummy variable. In this case the intercept could be interpreted as the expected effect for the second (or first) outcome type, and the regression weight of the dummy variable as the difference between the expected effects for both kinds of outcomes.

Table 1 shows the parameter estimates of this three-level mixed effects model without intercept (Eq. 10) as well as of the corresponding two-level mixed effects model ignoring the dependence (Eq. 11):

$$d_{jk} = \gamma_{100}X_{1jk} + \gamma_{200}X_{2jk} + v_{jk} + r_{jk} \tag{11}$$

Figure 1 presents the confidence intervals for the overall mean effects (Eqs. 6 and 9) and for the mean effects for both types of outcomes (Eqs. 10 and 11). Although the estimate of the mean effect for the second type of outcome is somewhat larger than for the first type of outcome, the difference between both types is statistically not significant, $p = .58$ and $p = .48$ for the two- and three-level models, respectively. The relatively small difference between the two types of outcomes is also illustrated by the fact that the variance estimate at the outcome level did not (visibly) change by including the type indicators in the model. The estimated percentage of estimated between-outcomes heterogeneity, $R^2_{(2)} = 100 * \left(1 - \frac{\widehat{\sigma}^2_{u(1)}}{\widehat{\sigma}^2_{u(0)}}\right)$ with $\widehat{\sigma}^2_{u(1)}$ and $\widehat{\sigma}^2_{u(1)}$ defined as the between-outcomes variance estimates for the model with and without the predictor, is zero. The variance at the study level even increases with the inclusion of the predictor, as sometimes happens in multilevel analyses while including predictors. Because $R^2$ is by definition positive, the resulting negative estimate of the percentage explained variance at the study level is truncated to zero (M.W.-L. Cheung, 2013).

Again the results illustrate that by considering effect sizes as dependent, larger standard errors and confidence intervals are obtained. Furthermore, the effect on the standard errors is much more pronounced for the second category of outcomes than for the first category. This can be explained by the fact that studies typically reported only one or a few outcomes of the first type (2.6 on average) but plenty of the second type (12.5 on average).

In the example, the estimate of the variance between studies (0.015) is relatively small compared to the estimated variance between outcomes within studies (0.046) and the sampling variance (the sampling variance depends on the size of the study, with a median estimate equal to 0.078). This means that for typical studies, the total variance estimate in observed effect sizes is equal to (0.078 +0.046 +0.015) =0.139, and that only about 11 % (0.015/0.139) of this total variance is variance between studies. The variance between outcomes from the same study is about 33 % of the total variance. These ratios are intraclass correlation coefficients that are often calculated in multilevel modelling (Raudenbush & Bryk, 2002), and correspond to the measures $I^2_{(3)}$ and $I^2_{(2)}$ that are proposed by M.W.-L. Cheung

[1] Three outlying standardized mean differences (>2) were not included in the analysis because of their substantial impact on the parameter estimates, especially on the variance estimates. The analysis therefore is based on 584 observed effect sizes from 39 studies.

**Table 1** Parameter estimates (and standard errors) from the child abuse and neglect meta-analysis

| Parameter | | Overall outcomes | | Accounting for outcome type | |
|---|---|---|---|---|---|
| | | Three-level (Eq. 6) | Two-level (Eq. 9) | Three-level (Eq. 10) | Two-level (Eq. 11) |
| Fixed coefficients | | | | | |
| Overall mean | $\gamma_{00}$ | 0.23 (0.026) | 0.20 (0.013) | | |
| Child abuse and neglect indicators | $\gamma_{100}$ | | | 0.20 (0.040) | 0.18 (0.035) |
| Risk factors | $\gamma_{200}$ | | | 0.23 (0.027) | 0.20 (0.014) |
| Variance components | | | | | |
| Between-study variance | $\sigma_u^2$ | 0.015 | | 0.016 | |
| (percentage of total variance) | | (11 %) | | (11 %) | |
| Between-outcome variance | $\sigma_v^2$ | 0.046 | 0.059 | 0.046 | 0.059 |
| (percentage of total variance) | | (33 %) | (43 %) | (33 %) | (43 %) |
| Median sampling variance | $\sigma_{r_{jk}}^2$ | 0.078 | 0.078 | 0.078 | 0.078 |
| (percentage of total variance) | | (56 %) | (57 %) | (56 %) | (57 %) |

*Note*: In all analyses, outlying values (observed effect sizes larger than 2) were excluded

(2013) in order to extend the measure $I^2$ commonly used in two-level meta-analyses to express the proportion of variance that is due to between study heterogeneity rather than to sampling variance. Because the null hypothesis of the homogeneity test for three-level models is the same as for two-level models, that is, the variance in observed effect sizes is due to sampling variance alone, Cochran's homogeneity test can be used without any adaptation (M.W.-L. Cheung, 2013). The test clearly shows that there is heterogeneity, Q =3196, df =583, $p < .0001$. As explained above, the variance at the study level can be interpreted as the overall covariance between outcomes. Dividing this covariance by the product of the standard deviations of the outcomes therefore yields an estimate of the between-outcome correlation. Assuming a common variance for two outcomes, 0.015/0.139 =0.11 thus refers to the correlation between a pair of outcomes. This relatively small intraclass correlation (here the correlation between outcomes within studies) also explains why in our example ignoring the dependence in effect sizes only makes a large difference if the number of outcomes per study is large.

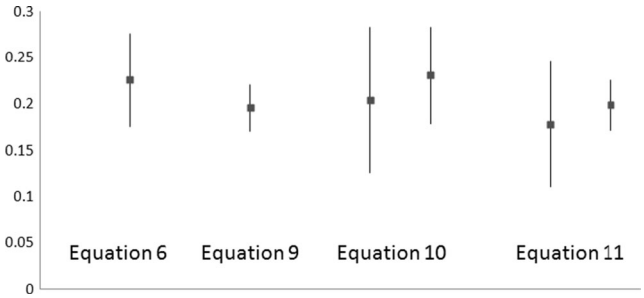


**Fig. 1** The 95 % confidence intervals for mean effect estimates for the child abuse and neglect meta-analysis. *Note*: Equations 6 and 10 are three-level models, Equations 9 and 11 two-level models. For Equations 9 and 10, the left confidence intervals refer to the mean effect on the child abuse and neglect indicators, the right ones to the mean effect on the risk factors

Another interesting result is that in general the mean effect estimates appeared to be larger when using the three-level modeling approach. This is due to a small negative correlation between the number of outcomes reported in the study and the average observed effect size in the study: if a study reports only a few outcomes, these reported effects are on average somewhat larger. The correlation was equal to -.09 and -.24 for the two types of outcomes, respectively. A possible explanation of this finding is reporting bias: in studies reporting the results on only a few outcome variables authors might have picked out the most remarkable findings, dropping the smaller or statistically non-significant findings. This kind of reporting bias will have a larger effect on the results if the dependence in outcomes is taken into account (this is also true for other approaches for dependent effect sizes), because in this case individual effect sizes from studies reporting more effect sizes are given relatively little weight.

## Simulation study

To evaluate the three-level approach for estimating the overall effect (over outcomes and studies), meta-analytic data sets were generated in two steps. In a first step, we generated raw data that could have been observed for a set of experimental studies with a control and an experimental group and continuous outcome variables. In the second step, the simulated raw data from each study were used to calculate a standardized mean difference, more specifically the difference between the experimental group mean and the control group mean, divided by the square root of the pooled within group variances. After generating meta-analytic datasets with multiple observed effect sizes per study, meta-analyses were performed using a two- and a three-level model, and results are compared.

To simulate the raw data in the first step, we used a multivariate two-level model, with a sample level and a between-study level (Kalaian & Raudenbush, 1996). At the first level, $Y_{ijk}$ is the score for study participant $i$ from study $k$ on outcome variable $j$ ($j$ =1, 2, …, $J$) and is regressed on a treatment dummy variable, equalling 1 for a participant belonging to the treatment group, and 0 for a participant from the control group.

$$\begin{cases} Y_{i1k} = \beta_{1k}(Treatment)_{ik} + e_{i1k} \quad \text{with} \quad \mathbf{e} \sim N(\mathbf{0}, \mathbf{V}) \\ \quad \vdots \\ Y_{iJk} = \beta_{Jk}(Treatment)_{ik} + e_{iJk} \end{cases} \tag{12}$$

The vector of residuals, $\boldsymbol{e}$, is assumed to follow a multivariate normal distribution with zero means, and a $J$ x $J$ covariance matrix $\mathbf{V}$. Therefore, the expected value in study $k$ for outcome $j$ is equal to 0 if the participant is measured in the control condition, and $\beta_{jk}$ if the participant is measured in the treatment condition. Because the interest in the meta-analysis is in the treatment effect (expressed as the standardized difference between the treatment and control group means) and not in the expected value in the control condition, we did not include an intercept in the model used to generate the data (otherwise stated, the intercept is equal to zero). This is not a restriction of our simulation though, because analyses will be done on standardized mean differences which are not affected by the value of the intercept. The effect of the treatment for each outcome possibly varies over studies. Therefore, we used an additional set of equations at the second level, the study level:

$$\begin{cases} \beta_{1k} = \gamma_{10} + w_{1k} \qquad \text{with} \quad \mathbf{w}_k \sim N(\mathbf{0}, \Omega_w) \\ \quad \vdots \\ \beta_{Jk} = \gamma_{J0} + w_{Jk} \end{cases} \tag{13}$$

in which $\gamma_{j0}$ ($j$ =1, 2, …, $J$) refers to the mean treatment effect for outcome $j$ in the population of studies, and $w_{jk}$ to the deviation of the treatment effect in study $k$ from this mean effect. As in the sample level equation, the vector of residuals, $\boldsymbol{w_k}$, is assumed to follow a multivariate normal distribution with zero means, and a $J$ x $J$ covariance matrix $\Omega_w$.

In the second step of our data generation, data for each outcome within each study were summarized using standardized mean differences. The expected value of the standardized mean difference for outcome $j$ in study $k$ is approximately equal to the population standardized mean difference, which is equal to $\frac{\beta_{jk}}{\sigma_e}$. The observed effect size nevertheless can deviate from this population value because both the mean difference and the residual standard deviation is estimated. A correction factor was used to correct these standardized mean differences for small-sample bias (Hedges, 1981): $d = \left(1 - \frac{3}{4(n_E+n_C)} - 9\right) \frac{\overline{Y}_E - \overline{Y}_C}{s_p}$. These standardized mean differences were analyzed using the three-level model described above (Eq. 6).

Data were generated under several conditions. We start by describing the simulation of balanced data. We refer to these conditions as the reference conditions, because after simulating and analyzing these balanced data, we relaxed one by one the strict assumptions and compared the results with the results for the balanced data. A summary of the conditions is given in Table 2.

## Reference conditions

We generated effect sizes for seven outcome variables in each study ($J$=7). The number of studies ($K$) was equal to 30 or 60, both group sizes within studies ($n$) equal to 25. The variance at the first level of our multivariate model (this is $\sigma_e^2$) was equal to 1 for all outcomes. We want to remark that using a single value for the level 1 variance is not a real restriction, because we study the meta-analysis of standardized effect sizes: if for a study all scores are multiplied by two, both the difference in mean and the level-1 standard deviation are multiplied by two, but the standardized mean difference remains unchanged. Therefore, we did not vary the level-1 variance over studies or outcomes, only the standardized mean difference. Because the choice of the value for the level-1 variance is trivial, we chose to use a value of 1 for all outcomes, so that $\beta_{jk}$ of Equation 12 can be interpreted as the population standardized mean difference for outcome $j$ in study $k$. We simulated data sets without systematic heterogeneity between outcomes ($\gamma_{10}$ =…= $\gamma_{70}$ =.40) and data with outcome-specific mean effects ($\gamma_{10}$=.10; $\gamma_{20}$=.20; $\gamma_{30}$=.30; $\gamma_{40}$=.40; $\gamma_{50}$=.50; $\gamma_{60}$=.60; $\gamma_{70}$=.70), values that were chosen to be representative for the small to large effects commonly found in empirical behavioral research (Cohen, 1988). We further varied the covariance between outcomes' raw data at the sample level ($\sigma_{e_j e_{j'}}$ = 0 or 0.40), having immediate implications for the covariance between outcomes' effect size data. Given that the first-level variance is equal to 1 for all outcomes, the covariance of 0.40 corresponds to a correlation of 0.40. In behavioral research, such a correlation can be regarded as a relatively large correlation, and therefore ensures that the simulated dependence is large enough to make its possible effect on inferences clearly visible. Equation 2 can be used to calculate the corresponding sampling covariance of the observed treatment effect sizes for two outcomes with mean effect 0.40. These values are 0 and 0.032 respectively. If the mean effect depends on the outcome, the effect size sampling covariance

**Table 2** Overview of the characteristics of the generated data

| | Sample | | | Outcomes | | Studies | | |
|---|---|---|---|---|---|---|---|---|
| | $N$ | $\sigma_e^2$ | $\sigma_{e_j e_{j'}}$ | $J$ | effects | $K$ | $\sigma_w^2$ | $\sigma_{w_{jk} w_{j'k}}$ |
| Reference conditions | 25 | 1 | 0 or 0.40 | 7 | Homogeneous or heterogeneous fixed | 30 or 60 | 0.10 | 0 or 0.04 |
| Varying number of outcomes | 25 | 1 | 0 or 0.40 | 1 - 7 | Homogeneous or heterogeneous fixed | 30 or 60 | 0.10 | 0 or 0.04 |
| Varying between-study variance | 25 | 1 | 0 or 0.40 | 7 | Homogeneous or heterogeneous fixed | 30 or 60 | 0.025, 0.10 & 0.40 | 0 or 0.04 |
| Varying intercorrelations | 25 | 1 | 0 or (0.20 & 0.80) | 7 | Homogeneous or heterogeneous fixed | 30 or 60 | 0.10 | 0 or 0.04 |
| Varying study sizes | $\mu$ =25 $\sigma$ =15.12 | 1 | 0 or 0.40 | 7 | Homogeneous or heterogeneous fixed | 30 or 60 | 0.10 | 0 or 0.04 |
| Randomly sampled outcomes | 25 | 1 | 0 or 0.40 | 7 | Homogeneous or heterogeneous random | 30 or 60 | 0.10 | 0 or 0.04 |
| Combination | $\mu$ =25 $\sigma$ =15.12 | 1 | 0 or (0.20 & 0.80) | 1-7 | Homogeneous or heterogeneous random | 30 or 60 | 0.025, 0.10 & 0.40 | 0 or 0.04 |

in case $\sigma_{e_j e_{j'}}$ = 0.40 for some pairs of outcomes can be slightly different from 0.032. We also varied the covariance in treatment effects between outcomes at the study level ($\sigma_{w_{jk} w_{j'k}}$ = 0 or 0.04). A positive covariance means that if in a study we find a relatively large effect for one outcome, we also expect a relatively large effect for another outcome (for instance because in this study the treatment was relatively intensive). The between-study variance in effect sizes, $\sigma_w^2$, was chosen to be 0.10. Given that for studies with $n$ =25 the sampling variance of observed effect sizes is about 0.08 (Eq. 1), the resulting ratio of the between-study variance and the sampling variance is realistic and avoids that the effects of correlation at either level become ignorable (Riley, 2009). Covariances at both levels were defined to represent a null versus a relatively large correlation in outcomes (covariances correspond with correlation coefficients of 0 and0.40). Large differences between the outcomes in the mean effects were chosen to guarantee that the effect of outcome heterogeneity on the results would be clearly visible.

These simulation conditions and analyses are similar to those in our previous study (Van den Noortgate et al., 2013). There are, however, differences in several aspects. First, the number of outcomes ($J$ =7) is larger than in our previous study ($J$ =2 or 5). Secondly, whereas in our previous study we especially reported the results of the raw data analyses, in this simulation study we only perform and report on effect size analyses. Third, whereas the number of conditions is equal to 16, this is 2 (number of studies) ×2 (heterogeneity over outcomes) ×2 (sampling covariance) ×2 (between-study covariance), in the previous simulation we varied more parameters, resulting in 432 conditions. We limited the number of conditions in this paper, because the focus now is on the effect of relaxing the strict assumptions underlying the reference conditions.

### Varying the number of outcomes per study

A first assumption in the reference conditions is that the number of outcomes is exactly the same in each study. In practice, however, the number of outcomes often varies over studies, as was the case in the meta-analysis of Geeraert et al. (2004). To investigate the effect of a varying number of outcomes per study on the results of the three-level analysis, we first simulated data sets in the same way as in the reference conditions, and in a second step randomly deleted part of the effect sizes. More specifically, each effect size had a probability of0.50 to be deleted. In this way, each study consisted of a random sample of size zero to seven outcomes from the fixed set of seven outcomes, with an expected number of 3.5 outcomes. If for a study none of the outcomes was retained, the meta-analytic data set in practice did not include 30 or 60 studies anymore. However, this factor hardly affects our results because for each study the probability of not observing any of the seven outcomes is very small and therefore the total number of studies was always equal or close to 30 or 60.

### Varying between-study variances

It is not unlikely that the between-study variance in the treatment effect depends on the outcome. Whereas in the reference conditions the between-study variance was equal to 0.10 for all seven outcomes, in this extension we simulated data sets where in each data set an outcome variable had an equal (one-third) probability of having a between-study variance of 0.025, 0.10 or 0.40 (resulting in a mean variance of 0.175). To this end, we multiplied the residuals $w_{jk}$ of Equation 13 for a specific outcome (over all studies) with 0.5, 1 or 2. As a consequence, the covariances at the second level also vary (from 0.01 to 0.16 with an average of 0.07 ). With this exception, data were simulated as in the reference conditions.

Varying intercorrelations

As explained above, the variance at the study level can be interpreted as the covariance between outcomes. Because the three-level model includes only one parameter for the between-study variance, the three-level analysis implicitly assumes that the intercorrelations are constant for all pairs of outcomes. However, if two outcome variables refer to two operationalizations of the same construct, the correlation is likely to be larger than when two outcome variables refer to two different constructs. We released the assumption of a common correlation, and simulated data with the following distribution for the sample level residuals:

$$\mathbf{e} \sim N(\mathbf{0}, V) \text{ with } V = \begin{bmatrix} 1 & & & & & & \\ .8 & 1 & & & & & \\ .8 & .8 & 1 & & & & \\ .8 & .8 & .8 & 1 & & & \\ .2 & .2 & .2 & .2 & 1 & & \\ .2 & .2 & .2 & .2 & .8 & 1 & \\ .2 & .2 & .2 & .2 & .8 & .8 & 1 \end{bmatrix}. \quad (14)$$

More specifically, we assumed two groups of outcomes (a group of four outcomes and a group of three outcomes) with high correlations within groups but small correlations between groups.

Varying study sample sizes

In meta-analysis it is very uncommon that all studies have the same sample size (Osburn & Callender, 1992). Studies with a smaller size result in less precise effect sizes estimates, and therefore are given less weight in the meta-analysis. An interesting question therefore is how the three-level approach would perform if group sizes vary over studies. To answer this question, we randomly sampled the group size within studies using a lognormal distribution as follows: $n \sim (6+\text{lnN}(2.7,0.7))$, and rounded the group sizes afterwards to integer values. The distribution has an expected value of 25 (and therefore results are comparable with the other simulation results, for which we used group sizes of 25 for all studies), has a standard deviation of 15.12, is positively skewed (as is often the case in meta-analytic data sets), and has a minimum value of 6. Subsequently, data were simulated as in the reference conditions.

Randomly sampling outcomes

Until now, we assumed that each of the studies uses (a subset of) the same seven outcome variables. A fixed set of outcomes can for instance be plausible in biomedical studies in which the systolic and diastolic blood pressure might be commonly used criteria to evaluate some medication. In the example of the meta-analysis of Geeraert et al. (2004), however, we saw that there is not always consensus about the outcome variables to be used, and each study may use its own idiosyncratic measures. To simulate this situation, we assumed that in each study seven outcome variables are used, but those variables were randomly sampled from a population of possible outcomes. More specifically, seven mean effects were sampled for each study from a normal distribution, $N(.40,.04)$. Because as before the population effects of the studies are assumed to randomly vary around the mean effects for the outcome variables that are measured in that study, $w_{jk} \sim N(0,.10)$, and these residuals are independent of the size of the mean outcome effect, $\beta_{jk} \sim N(.40,.14)$. For the conditions without heterogeneity in outcome effects, we did not simulate and analyze new data, because the situation is the same as in the reference conditions for a common outcome effect.

Combination

In practice, a combination of these extensions can be expected. Therefore, we simulated data combining all these extensions: for each study, we assumed a random sample of outcome variables, the number of outcomes and the group sizes varied over studies, and the between-study variance and the intercorrelations varied over outcomes.

In sum, we explored the so called reference conditions and six extensions. Each time we had 16 conditions (two study sizes, two values for the within-study covariance, two values for the between-study covariance, heterogeneity or no heterogeneity over outcomes). Only in the random sample extension we did not generate new data for the no-heterogeneity case, as explained above. Therefore, in total we have 104 conditions, $(6 \times 16) + (1 \times 8)$. For each combination, we simulated 1,000 meta-analytic data sets, 104,000 in total. Given the large number of conditions, and given that we do not want to draw conclusions about specific conditions, but rather about trends over conditions, 1,000 datasets per condition can be regarded as sufficiently large. Each data set was analyzed twice: using the three-level model (Eq. 6) as well as using an ordinary random effects model, ignoring the dependence in effect sizes (Eq. 9). Compared to the three-level model, the model ignores the highest level, the study level.

To evaluate the use of the two- and three-level model approach, we will look at the estimates of the mean effect, of the corresponding standard error, and of the variance components. By definition, the standard error for the mean effect parameter is equal to the standard deviation of the sampling distribution of the mean effect estimator. Because in each condition we have a relatively large number of data sets (1,000), the standard deviation of the resulting 1,000 estimates of the mean effect can be considered as a good approximation of the standard deviation of the sampling distribution.

Therefore, we will compare the mean (estimated) standard errors with the standard deviation of the estimates of the mean effect. A second way to evaluate the standard errors is to investigate the coverage proportion of the confidence intervals. For each condition we will calculate for each data set a 90 % and a 95 % confidence interval around the mean effect estimate using the standard error estimates. We will evaluate for each condition what proportion of the 1,000 confidence intervals included the true overall mean effect (equal to 0.40 for all conditions). When confidence intervals are accurate, this proportion should by definition be close to 0.90 or 0.95, respectively.

## Results of the Simulation Study

Because we found that the bias in estimating the mean effect was relatively small in all conditions, in the remainder of the article we will focus on the standard errors and variance component estimates.

### Reference conditions

The upper left part of Table 3 gives the mean variance estimates for the balanced data, for the case where the mean population effect is the same for all seven outcomes and the number of studies is 30. Table 4 gives the variance estimates if the number of studies is 60. The results show that if a traditional two-level meta-analytic model (ignoring the dependence) is used, the estimated variance between outcomes (around 0.09) is somewhat smaller than the between-study variance for each outcome variable from the multivariate model we used to generate the data (0.10). We found that if the raw data are analyzed directly (results not shown here) using a two-level model, the mean estimate of the between-outcome variance is exactly equal to 0.10. This negative bias in the estimate of the systematic variance when analyzing the effect sizes is very similar to the bias observed in earlier simulation studies for the meta-analysis of effect sizes for the same size and number of studies. For an explanation of this bias, we refer to Equation 1, expressing the sampling variance of the standardized mean differences. The sampling variance for each observed effect size is estimated by replacing the unknown delta in the formula by the observed effect size. As explained above, effect sizes with larger estimated sampling variance are given less weight in the meta-analysis. Because larger observed effect sizes that are further from zero will receive smaller weights than relatively small observed effect sizes, the variance will be underestimated. This is especially true when studies are relatively small (as in our simulation study), because in this case observed effect sizes will vary a lot and weights depend to a large extent on these observed effect sizes.

If a three-level model is used, this variance is distributed over the study level and the outcome (within-study) level. Earlier we showed that in the three-level model, the variance between studies can be interpreted as the covariation between outcomes from the same study (Eq. 8). This is confirmed by our simulation study: the estimated between-study variance is on average close to the total covariance between two outcomes from the same study, this is the covariance between two outcomes in population effects, plus the sampling covariance as calculated using Equation 2. More specifically, the expected covariance is 0 (if $\sigma_{e_j e_{j'}} = 0$ and $\sigma_{w_j w_{j'}} = 0$ ), 0.04 (if $\sigma_{e_j e_{j'}} = 0$ and $\sigma_{w_j w_{j'}} = 0.04$), 0.032 (if $\sigma_{e_j e_{j'}} = 0.4$ and $\sigma_{w_j w_{j'}} = 0$ ), or 0.072 (if $\sigma_{e_j e_{j'}} = 0.4$ and $\sigma_{w_j w_{j'}} = 0.04$ ). The remainder of the total variance is attributed to the outcome level within studies.

If the mean effect size depends on the outcome (upper right part of Tables 3 and 4), the total variance is larger. More specifically, the variance of the mean outcome effect values used to simulate data (the variance of 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, which is equal to 0.04) is added to the total variance: whereas previously the total estimated variance for the two- and three-level model was around 0.09, the total estimated variance now is around 0.13. For the three-level model, we see that the between-study variance is slightly decreased, but that the between-outcome variance is increased.

Despite the negative bias in variance estimates, mean standard errors of the three-level model are quite accurate: they closely resemble the standard deviations of the mean effect size estimates, as shown in the upper part of Table 5 (for $K$ =30) and Table 6 (for $K$ =60). Only if the (mean) effect depends on the outcome (upper right) and there is no correlation between effect sizes at either level, the standard errors are on average slightly too large: whereas the standard deviations of the estimates are not affected by the heterogeneity of the effects over outcomes, the mean standard error is slightly larger. As expected, the standard deviation (and standard error) is getting larger with an increasing within-study covariance and with an increasing between-study covariance between outcomes. That the fluctuation in estimates is larger with an increasing covariance, is because the higher the covariation, the larger the overlap in information given by the observed effect sizes, and therefore the smaller the information on the overall effect size that is given by the set of observed effect sizes.

Regarding the two-level model, we found that the standard deviations of the mean effect estimates are almost identical compared to those of the three-level model. However, the standard errors clearly are not affected by a possible correlation between the outcomes. Therefore, standard errors were found (more or less) appropriate only if there is no correlation at either level, too small otherwise.

**Table 3** Variance estimates multiplied by 1,000 for the three- and two-level meta-analytic models (K =30)

| | | | Common effect | | | Outcome-specific effect | | |
|---|---|---|---|---|---|---|---|---|
| | | | Three-level | | Two-level | Three-level | | Two-level |
| | $\sigma_{e_j e_{j'}}$ | $\sigma_{w_j w_{j'}}$ | $\widehat{\sigma}_u^2$ | $\widehat{\sigma}_v^2$ | $\widehat{\sigma}_v^2$ | $\widehat{\sigma}_u^2$ | $\widehat{\sigma}_v^2$ | $\widehat{\sigma}_v^2$ |
| Reference conditions | 0 | 0 | 0 | 89 | 91 | 0 | 128 | 129 |
| | | 0.04 | 38 | 54 | 91 | 31 | 98 | 129 |
| | 0.4 | 0 | 30 | 60 | 90 | 24 | 105 | 129 |
| | | 0.04 | 69 | 23 | 90 | 62 | 68 | 130 |
| Varying number of outcomes | 0 | 0 | 1 | 85 | 90 | 0 | 128 | 130 |
| | | 0.04 | 34 | 53 | 90 | 26 | 102 | 130 |
| | 0.4 | 0 | 27 | 58 | 86 | 22 | 106 | 131 |
| | | 0.04 | 68 | 21 | 87 | 59 | 70 | 131 |
| Varying between-study variance | 0 | 0 | 0 | 140 | 144 | 0 | 178 | 180 |
| | | 0.04 | 44 | 97 | 140 | 40 | 144 | 181 |
| | 0.4 | 0 | 31 | 111 | 143 | 24 | 157 | 183 |
| | | 0.04 | 76 | 65 | 142 | 70 | 111 | 179 |
| Varying intercorrelations | 0 | 0 | 0 | 89 | 91 | 0 | 128 | 129 |
| | | 0.04 | 38 | 54 | 91 | 31 | 98 | 129 |
| | 0.46 | 0 | 34 | 57 | 92 | 28 | 101 | 129 |
| | | 0.04 | 71 | 19 | 89 | 66 | 63 | 127 |
| Varying study sizes | 0 | 0 | 0 | 89 | 91 | 0 | 128 | 129 |
| | | 0.04 | 37 | 54 | 92 | 31 | 98 | 129 |
| | 0.4 | 0 | 34 | 61 | 90 | 28 | 104 | 129 |
| | | 0.04 | 71 | 26 | 87 | 66 | 67 | 126 |
| Randomly sampled outcomes | 0 | 0 | 0 | 89 | 91 | 2 | 124 | 129 |
| | | 0.04 | 38 | 54 | 91 | 41 | 89 | 129 |
| | 0.4 | 0 | 30 | 60 | 90 | 32 | 97 | 128 |
| | | 0.04 | 69 | 23 | 90 | 72 | 59 | 128 |
| Combination | 0 | 0 | 0 | 149 | 155 | 0 | 186 | 196 |
| | | 0.04 | 48 | 108 | 154 | 48 | 142 | 193 |
| | 0.46 | 0 | 38 | 118 | 155 | 39 | 156 | 195 |
| | | 0.04 | 88 | 70 | 152 | 90 | 107 | 191 |

*Note*: $\sigma_{e_j e_{j'}}$ = level 1 (within-study) covariance between outcomes' raw data; $\sigma_{w_j w_{j'}}$ = level 2 (between-study) covariance between outcomes' treatment effects (as defined in the multivariate model, Equations 12-13), $\widehat{\sigma}_u^2$ = between-study variance estimate of the three-level model (Eq. 6) and $\widehat{\sigma}_v^2$ the between-outcome variance estimate of the three-level model (Eq. 6) and two-level model (Eq. 9)

The findings regarding the standard errors help us in understanding the coverage proportions (CP) of the 90 and 95 % confidence intervals around the mean effect estimate. The upper part of Table 7 (for $K$ =30; very similar results were found for $K$ =60) shows that for the three-level model, the CP is quite accurate in all cases, although somewhat too large if effects are outcome-specific and if there is no covariance at either level. For the traditional random effects model, the two-level model, the CP is only approximately equal to the nominal level if there is no covariance at either level. If outcomes covary, the CP can become much too small, with CPs for the 95 % confidence intervals going down till 71 %.

Varying the number of outcomes per study

Comparing the first two blocks of Tables 3 and 4 reveals that if in studies only a subset of the seven outcome variables is measured, variance parameter estimates are hardly affected if $K$ =30: the means of the 1,000 estimates for each combination are similar to those of the reference conditions, especially when $K$ =60. Standard errors are however larger than in the reference conditions (Tables 5 and 6), which can be explained by the smaller total number of effect sizes that is used to estimate the mean effect size: the expected number of effect sizes per study equals 3.5 instead of 7. This effect is

**Table 4** Variance estimates multiplied by 1,000 for the three- and two-level meta-analytic models (K =60)

| | | | Common effect | | | Outcome-specific effect | | |
|---|---|---|---|---|---|---|---|---|
| | | | Three-level | | Two-level | Three-level | | Two-level |
| | $\sigma_{e_j e_{j'}}$ | $\sigma_{w_j w_{j'}}$ | $\widehat{\sigma}_u^2$ | $\widehat{\sigma}_v^2$ | $\widehat{\sigma}_v^2$ | $\widehat{\sigma}_u^2$ | $\widehat{\sigma}_v^2$ | $\widehat{\sigma}_v^2$ |
| Reference conditions | 0 | 0 | 0 | 91 | 93 | 0 | 130 | 131 |
| | | 0.04 | 37 | 54 | 91 | 31 | 100 | 132 |
| | 0.4 | 0 | 30 | 61 | 91 | 24 | 106 | 130 |
| | | 0.04 | 68 | 23 | 90 | 61 | 68 | 129 |
| Varying number of outcomes | 0 | 0 | 0 | 89 | 93 | 0 | 129 | 130 |
| | | 0.04 | 37 | 53 | 90 | 30 | 101 | 132 |
| | 0.4 | 0 | 30 | 60 | 91 | 23 | 109 | 133 |
| | | 0.04 | 68 | 22 | 90 | 59 | 70 | 131 |
| Varying between-study variance | 0 | 0 | 0 | 149 | 152 | 0 | 191 | 192 |
| | | 0.04 | 48 | 103 | 149 | 41 | 149 | 189 |
| | 0.4 | 0 | 30 | 122 | 153 | 23 | 168 | 192 |
| | | 0.04 | 77 | 71 | 149 | 71 | 118 | 189 |
| Varying intercorrelations | 0 | 0 | 0 | 91 | 93 | 0 | 130 | 131 |
| | | 0.04 | 37 | 54 | 91 | 31 | 100 | 132 |
| | 0.46 | 0 | 35 | 57 | 91 | 29 | 101 | 129 |
| | | 0.04 | 74 | 18 | 91 | 68 | 63 | 130 |
| Varying study sizes | 0 | 0 | 0 | 89 | 91 | 0 | 129 | 129 |
| | | 0.04 | 38 | 54 | 91 | 31 | 97 | 128 |
| | 0.4 | 0 | 34 | 62 | 90 | 27 | 105 | 128 |
| | | 0.04 | 73 | 27 | 90 | 65 | 68 | 127 |
| Randomly sampled outcomes | 0 | 0 | 0 | 91 | 93 | 0 | 129 | 131 |
| | | 0.04 | 37 | 54 | 91 | 38 | 92 | 130 |
| | 0.4 | 0 | 30 | 61 | 91 | 30 | 99 | 130 |
| | | 0.04 | 68 | 23 | 90 | 68 | 61 | 128 |
| Combination | 0 | 0 | 0 | 152 | 157 | 0 | 190 | 196 |
| | | 0.04 | 49 | 105 | 154 | 49 | 145 | 196 |
| | 0.46 | 0 | 37 | 121 | 157 | 38 | 159 | 196 |
| | | 0.04 | 87 | 71 | 153 | 89 | 109 | 195 |

*Note*: $\sigma_{e_j e_{j'}}$ = level 1 (within-study) covariance between outcomes' raw data; $\sigma_{w_j w_{j'}}$ = level 2 (between-study) covariance between outcomes' treatment effects (as defined in the multivariate model, Equations 12-13), $\widehat{\sigma}_u^2$ = between-study variance estimate of the three-level model (Eq. 6) and $\widehat{\sigma}_v^2$ the between-outcome variance estimate of the three-level model (Eq. 6) and two-level model (Eq. 9)

especially clear when there is no correlation between outcomes, because in this case each effect size gives information that does not overlap with information given by the other effect sizes of the study. Table 7 shows similar CPs as those for the reference conditions. Although the two-level model still yields unacceptable CPs, it performs better than in the reference condition. This can again be explained by the decreased expected number of outcomes per study: the more effect sizes are reported per study, the more problematic is the ignorance of the overlap in information.

### Varying between-study variances

Tables 3 and 4 reveal that if the effect is common to all seven outcome variables, the total variance is larger than in the reference conditions, more specifically between 0.140 and 1.50. This is not unexpected because now the values for the between-study variance that were used to generate the data are on average equal to 0.175 rather than 0.100. This also means we observe again a negative bias in the total variance estimate. This variance is again split into two parts when using a three-level model. The part that is going to the study level is zero if

**Table 5** Mean standard error estimate (SE) multiplied by 1,000 for the mean effect estimate and standard deviation (STDEV) of the mean effect estimates (K =30)

| | | | Common effect | | | | Outcome-specific effect | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Three-level | | Two-level | | Three-level | | Two-level | |
| | $\sigma_{e_j e_{j'}}$ | $\sigma_{w_j w_{j'}}$ | SE | STDEV | SE | STDEV | SE | STDEV | SE | STDEV |
| Reference conditions | 0 | 0 | 30 | 28 | 29 | 28 | 32 | 28 | 32 | 28 |
| | | 0.04 | 44 | 43 | 29 | 43 | 44 | 43 | 32 | 43 |
| | 0.4 | 0 | 41 | 43 | 29 | 43 | 41 | 42 | 32 | 42 |
| | | 0.04 | 53 | 54 | 29 | 54 | 53 | 52 | 32 | 52 |
| Varying number of outcomes | 0 | 0 | 41 | 37 | 40 | 37 | 45 | 38 | 44 | 38 |
| | | 0.04 | 50 | 51 | 40 | 52 | 51 | 52 | 44 | 52 |
| | 0.4 | 0 | 47 | 46 | 39 | 46 | 50 | 46 | 44 | 47 |
| | | 0.04 | 57 | 55 | 39 | 57 | 60 | 55 | 44 | 57 |
| Varying between-study variance | 0 | 0 | 34 | 33 | 33 | 33 | 36 | 34 | 35 | 34 |
| | | 0.04 | 48 | 51 | 33 | 51 | 49 | 50 | 36 | 50 |
| | 0.4 | 0 | 44 | 44 | 33 | 44 | 44 | 45 | 36 | 45 |
| | | 0.04 | 57 | 57 | 33 | 57 | 57 | 56 | 35 | 56 |
| Varying intercorrelations | 0 | 0 | 30 | 28 | 29 | 28 | 32 | 28 | 32 | 28 |
| | | 0.04 | 44 | 43 | 29 | 43 | 44 | 43 | 32 | 43 |
| | 0.46 | 0 | 43 | 44 | 29 | 44 | 43 | 43 | 32 | 43 |
| | | 0.04 | 54 | 53 | 29 | 53 | 54 | 54 | 32 | 54 |
| Varying study sizes | 0 | 0 | 31 | 30 | 30 | 30 | 33 | 29 | 33 | 29 |
| | | 0.04 | 45 | 44 | 30 | 44 | 44 | 43 | 33 | 43 |
| | 0.4 | 0 | 44 | 42 | 30 | 41 | 43 | 43 | 33 | 42 |
| | | 0.04 | 55 | 57 | 30 | 56 | 55 | 55 | 33 | 55 |
| Randomly sampled outcomes | 0 | 0 | 30 | 28 | 29 | 28 | 33 | 33 | 32 | 33 |
| | | 0.04 | 44 | 43 | 29 | 43 | 46 | 48 | 32 | 48 |
| | 0.4 | 0 | 41 | 43 | 29 | 43 | 44 | 44 | 32 | 44 |
| | | 0.04 | 53 | 54 | 29 | 54 | 55 | 55 | 32 | 55 |
| Combination | 0 | 0 | 50 | 48 | 48 | 48 | 54 | 51 | 52 | 50 |
| | | 0.04 | 60 | 60 | 48 | 60 | 63 | 62 | 52 | 63 |
| | 0.46 | 0 | 58 | 60 | 48 | 60 | 62 | 61 | 52 | 61 |
| | | 0.04 | 68 | 67 | 48 | 68 | 71 | 70 | 51 | 71 |

*Note*: $\sigma_{e_j e_{j'}}$ = level 1 (within-study) covariance between outcomes' raw data; $\sigma_{w_j w_{j'}}$ = level 2 (between-study) covariance between outcomes' treatment effects (as defined in the multivariate model, Eqs. 12 and 13)

there is no correlation between outcomes at either level, and is augmented with almost 0.032 if the within-study covariance is equal to 0.40. Although we would expect an increase of this variance with 0.04 if $\sigma_{w_j w_{j'}} = 0.04$, we found on average a slightly larger increase (about 0.045). For the outcome-specific case, the pattern is as expected, with the total variance being about 0.04 points higher than for the common-effect case.

The pattern for the standard errors (Tables 5 and 6) is very similar to the one for the reference conditions, with the exception that the standard errors are slightly larger, referring to the increased uncertainty due to the larger (mean) between-study variance. CPs are very similar to those of the reference conditions (Table 7).

### Varying intercorrelations

In the sampling covariance matrix given in Equation 14, the covariance between nine pairs of outcomes is 0.80, for twelve pairs the covariance is 0.20. On average, the covariance therefore is equal to 0.46. Extrapolating the effect of the sampling covariance that we found in the

**Table 6** Mean standard error estimate (SE) multiplied by 1,000 for the mean effect estimate and standard deviation (STDEV) of the mean effect estimates (K =60)

| | | | Common effect | | | | Outcome-specific effect | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Three-level | | Two-level | | Three-level | | Two-level | |
| | $\sigma_{e_j e_{j'}}$ | $\sigma_{w_j w_{j'}}$ | SE | STDEV | SE | STDEV | SE | STDEV | SE | STDEV |
| Reference conditions | 0 | 0 | 21 | 20 | 20 | 20 | 23 | 20 | 23 | 20 |
| | | 0.04 | 31 | 31 | 20 | 31 | 31 | 31 | 23 | 31 |
| | 0.4 | 0 | 29 | 28 | 20 | 28 | 29 | 29 | 23 | 29 |
| | | 0.04 | 37 | 37 | 20 | 37 | 37 | 36 | 22 | 36 |
| Varying number of outcomes | 0 | 0 | 30 | 29 | 29 | 29 | 32 | 29 | 32 | 29 |
| | | 0.04 | 36 | 36 | 29 | 36 | 38 | 37 | 32 | 37 |
| | 0.4 | 0 | 35 | 35 | 29 | 35 | 37 | 35 | 32 | 35 |
| | | 0.04 | 41 | 42 | 29 | 43 | 42 | 41 | 32 | 42 |
| Varying between-study variance | 0 | 0 | 24 | 24 | 24 | 24 | 26 | 24 | 26 | 24 |
| | | 0.04 | 35 | 35 | 24 | 35 | 35 | 36 | 26 | 36 |
| | 0.4 | 0 | 32 | 32 | 24 | 32 | 31 | 31 | 26 | 31 |
| | | 0.04 | 41 | 43 | 24 | 43 | 41 | 43 | 26 | 43 |
| Varying intercorrelations | 0 | 0 | 21 | 20 | 20 | 20 | 23 | 20 | 23 | 20 |
| | | 0.04 | 31 | 31 | 20 | 31 | 31 | 31 | 23 | 31 |
| | 0.46 | 0 | 30 | 30 | 20 | 30 | 30 | 30 | 23 | 30 |
| | | 0.04 | 38 | 39 | 20 | 39 | 38 | 38 | 23 | 38 |
| Varying study sizes | 0 | 0 | 22 | 21 | 21 | 21 | 23 | 20 | 23 | 20 |
| | | 0.04 | 32 | 31 | 21 | 32 | 31 | 31 | 23 | 31 |
| | 0.4 | 0 | 31 | 32 | 21 | 31 | 31 | 31 | 23 | 31 |
| | | 0.04 | 39 | 40 | 21 | 40 | 38 | 39 | 23 | 39 |
| Randomly sampled outcomes | 0 | 0 | 21 | 20 | 20 | 20 | 23 | 23 | 23 | 23 |
| | | 0.04 | 31 | 31 | 20 | 31 | 32 | 32 | 23 | 32 |
| | 0.4 | 0 | 29 | 28 | 20 | 28 | 31 | 30 | 23 | 30 |
| | | 0.04 | 37 | 37 | 20 | 37 | 38 | 39 | 22 | 39 |
| Combination | 0 | 0 | 36 | 34 | 35 | 34 | 38 | 38 | 37 | 37 |
| | | 0.04 | 43 | 41 | 35 | 42 | 46 | 46 | 37 | 46 |
| | 0.46 | 0 | 42 | 43 | 35 | 43 | 44 | 44 | 37 | 45 |
| | | 0.04 | 49 | 48 | 35 | 49 | 51 | 51 | 37 | 52 |

*Note*: $\sigma_{e_j e_{j'}}$ = level 1 (within-study) covariance between outcomes' raw data; $\sigma_{w_j w_{j'}}$ = level 2 (between-study) covariance between outcomes' treatment effects (as defined in the multivariate model, Eqs. 12 and 13)

reference conditions to a value of 0.46 would give almost identical results as those found for the conditions with these varying intercorrelations (Tables 3 and 4). We conclude that the between-study variance estimate of the three-level model refers to the mean covariance between outcomes. The same is true for the standard errors. Tables 5 and 6 reveal that the standard errors for the three-level model are equally appropriate for the conditions with a varying within-study covariance as compared to the conditions with a common covariance.

As a result, also the CPs are very similar (Table 7). We conclude that the three-level model is still appropriate when the correlations between outcomes are not constant.

### Varying study sizes

Variance estimates are very similar if the size of the study varies over studies as compared to the reference conditions with a common study size (Tables 3 and 4). Standard errors are in general only slightly higher, but still closely match the standard deviation of the estimates for the three-level model (Tables 5 and 6). Also the CPs are hardly affected when study sizes are variable rather than constant over studies (Table 7).

### Randomly sampling outcomes

The total variance for the heterogeneous case is underestimated to the same degree as in the reference conditions. The

**Table 7** Coverage percentages of the 90 % and 95 % confidence intervals for the three- and two-level meta-analytic models (K =30)

| | | | Common effect | | | | Outcome-specific effect | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Three-level | | Two-level | | Three-level | | Two-level | |
| | $\sigma_{e_j e_{j'}}$ | $\sigma_{w_j w_{j'}}$ | 90% | 95% | 90% | 95% | 90% | 95% | 90% | 95% |
| Reference conditions | 0 | 0 | 91 | 96 | 89 | 95 | 95 | 97 | 93 | 96 |
| | | 0.04 | 91 | 95 | 75 | 83 | 90 | 95 | 80 | 85 |
| | 0.4 | 0 | 89 | 93 | 72 | 80 | 90 | 94 | 78 | 84 |
| | | 0.04 | 90 | 96 | 62 | 71 | 90 | 96 | 69 | 77 |
| Varying number of outcomes | 0 | 0 | 94 | 98 | 93 | 97 | 96 | 99 | 96 | 98 |
| | | 0.04 | 90 | 95 | 79 | 88 | 90 | 95 | 84 | 90 |
| | 0.4 | 0 | 91 | 95 | 83 | 90 | 93 | 96 | 88 | 93 |
| | | 0.04 | 90 | 94 | 74 | 82 | 91 | 96 | 79 | 88 |
| Varying between-study variance | 0 | 0 | 92 | 96 | 91 | 94 | 92 | 96 | 91 | 95 |
| | | 0.04 | 88 | 94 | 70 | 79 | 90 | 94 | 78 | 85 |
| | 0.4 | 0 | 90 | 94 | 78 | 87 | 89 | 94 | 81 | 88 |
| | | 0.04 | 91 | 96 | 65 | 74 | 92 | 96 | 70 | 79 |
| Varying intercorrelations | 0 | 0 | 91 | 96 | 89 | 95 | 95 | 97 | 93 | 96 |
| | | 0.04 | 91 | 95 | 75 | 83 | 90 | 95 | 80 | 85 |
| | 0.46 | 0 | 89 | 95 | 71 | 80 | 89 | 95 | 76 | 84 |
| | | 0.04 | 90 | 95 | 61 | 69 | 90 | 95 | 66 | 74 |
| Varying study sizes | 0 | 0 | 91 | 96 | 88 | 94 | 95 | 98 | 93 | 98 |
| | | 0.04 | 90 | 95 | 72 | 81 | 91 | 95 | 79 | 87 |
| | 0.4 | 0 | 91 | 96 | 77 | 85 | 90 | 95 | 80 | 86 |
| | | 0.04 | 87 | 94 | 60 | 68 | 88 | 95 | 66 | 75 |
| Randomly sampled outcomes | 0 | 0 | 91 | 96 | 89 | 95 | 89 | 95 | 86 | 92 |
| | | 0.04 | 91 | 95 | 75 | 83 | 90 | 95 | 70 | 78 |
| | 0.4 | 0 | 89 | 93 | 72 | 80 | 91 | 96 | 78 | 86 |
| | | 0.04 | 90 | 96 | 62 | 71 | 89 | 95 | 64 | 71 |
| Combination | 0 | 0 | 92 | 97 | 89 | 95 | 92 | 97 | 90 | 95 |
| | | 0.04 | 90 | 95 | 82 | 88 | 90 | 95 | 82 | 88 |
| | 0.46 | 0 | 89 | 94 | 81 | 87 | 89 | 95 | 84 | 89 |
| | | 0.04 | 90 | 95 | 75 | 83 | 90 | 96 | 75 | 84 |

*Note*: $\sigma_{e_j e_{j'}}$ = level 1 (within-study) covariance between outcomes' raw data; $\sigma_{w_j w_{j'}}$ = level 2 (between-study) covariance between outcomes' treatment effects (as defined in the multivariate model, Equations 12-13). The standard error of the percentages in the cells is about 0.7 % for the 95 % CIs, and 0.9 % for the 90 % CIs

between-study variance is, however, slightly larger, and the between-outcome variance therefore is slightly smaller (Tables 3 and 4). This means that according to the results, the outcomes are slightly more correlated.

Compared to the reference conditions, the standard errors are very slightly higher. The major finding, however, is that the standard deviation of the estimates is now somewhat larger, especially if outcomes are not related. Mean standard errors now are almost identical to the standard deviations for all conditions. Hence, CPs for the three-level model are very close to the nominal levels (Table 7).

### Combination

In this last set of conditions, we relax all the assumptions at the same time. Differences with the reference conditions in the variance estimates (Tables 3 and 4) can be explained by combining the effects of each of the extensions separately. For instance, because the mean between-study variance is larger, the total variance for both the two-level and the three-level model is larger. Also the results for the standard errors (Tables 5 and 6) are as expected, with larger standard errors because of the larger mean between-study variance and the smaller number of observed effect sizes. Especially important is the finding that, as in the random sampling extension, the

standard errors of the three-level model are very accurate in all conditions. As a result, CPs are accurate for the three-level model for all conditions (Table 7). For the traditional two-level model, standard errors and CPs are again only accurate if there is no covariance at either level.

## Conclusions

In this paper, we evaluated the performance of a three-level approach to account for the dependence between multiple effect sizes within studies. The use of a three-level model to account for the sampling covariance between multiple outcomes per study is an appealing approach, as illustrated by Geeraert et al. (2004) in a real data meta-analysis. Probably the most important advantage of the multilevel approach for dealing with dependent effect sizes is that it does not require 'known' (or previously estimated) sampling covariances before performing the meta-analysis, as is required in a multivariate analysis. This is a major help to meta-analysts because information about the covariance between outcomes is only rarely reported in the primary studies or in other literature, especially if outcome variables are not measured using very common and well known scales.

Another advantage of the use of a three-level model is that it is a very flexible model that can be extended for instance by including characteristics of outcomes and studies as predictors, possibly explaining (part) of the variance at the outcome and study level. For instance, if a meta-analyst is not satisfied by merely estimating the overall effect and the variation in and over studies, but also wants an estimate of the mean effect for each separate outcome variable, the outcome level equation (Equation 4) of the three-level model can be adapted by dropping the intercept and regressing the observed effect sizes on a set of dummy variables, one for each type of outcome. The coefficients of these dummy variables then refer to the population effects for the separate outcomes. In our simulations, we found accurate standard errors and confidence proportions for models including dummy outcome indicators (not discussed or shown in this paper) for all extensions with the fixed set of seven outcomes. Although including a dummy variable for each outcome does not make sense if each study uses its own (random set of) outcomes, it might be possible to divide outcomes in broad categories, and including outcome category dummy variables in the model to obtain estimates of the mean effect for each outcome type, as was done in the example. Besides exploring differences between (types of) outcomes in the overall effect, interaction terms of the dummy variables and potential moderator variables can be included in the model to investigate differential moderator effects. By allowing the coefficients of dummy indicators for types of outcomes vary randomly at the second or third level, it is possible to model scenarios where the variance and covariance of effect sizes depend on the type of outcome. The multilevel approach can also be used to model other dependencies at the same time, for example, by defining an additional upper level of research team or countries that group the studies (Konstantopoulos, 2011).

The use of a three-level model does not require that the number of effect sizes is the same for all studies. It does not even require that all studies report more than one effect size. In this way, the multilevel meta-analysis makes efficient use of all available effect sizes. If at least one study reports more than one effect size, we can disentangle the between-study and between-outcome variance, but the accuracy of the between-outcome variance will depend on the average number of observed effect sizes per study (Maas & Hox, 2005), and a small total number of outcomes can result in unstable and even negative variance estimates at level 2 and level 3. In our previous simulation study (Van den Noortgate et al., 2013), we found that the three-level approach works well with only two outcomes per study, but more research is needed on the situation where most studies only report one effect size.

Moreover, all analyses can straightforwardly be performed using statistical software such as SAS or R (using the metaSEM package of M.W.-L. Cheung, 2013), or using multilevel software such as MLwiN or HLM, without requiring additional calculations. Excellent handbooks about multilevel analysis, including a discussion of multilevel meta-analysis, are available (e.g., Hox, 2002; Raudenbush & Bryk, 2002).

There are, however, limitations of the three-level approach, that are due to considering outcomes within a study as a random sample of possible outcomes, in the same way as in an ordinary random effects meta-analysis studies are regarded as a random sample of studies. A first limitation is that because conclusions are drawn on the population of outcomes rather than on the individual outcomes, the three-level model is especially appropriate in the (common) case where the meta-analyst is not primarily interested in the individual outcome and study effects, but rather aims at generalizing the results to a population of effects.

A second limitation of the three-level approach compared to the multivariate approach, is that it will not allow us to obtain joint confidence intervals for two outcomes, or to estimate the (between-study) correlation between a pair of outcomes, while this correlation might be of major interest in some applications (such as for estimating the trade off of specificity and sensitivity in a diagnostic test meta-analysis; Jackson et al., 2011).

A third limitation of the use of a three-level approach to estimate the average effect size, is that if a specific (type of) outcome is reported more frequently, this outcome will also have more weight in the estimation of the mean effect. Moreover, if some studies will report the observed effects

for outcomes for which the population effect is relatively small, while other studies will rather report results for outcomes with large effects, this will increase the between-study variance rather than the between-outcome variance. The three-level model indeed assumes that the outcomes reported in a study form a random sample from a population of outcomes.

A fourth limitation is that although the three-level model is intuitively appealing (more specifically the idea that there is not only variation between studies, but also between multiple effects within studies), the model is more complex than the model we would use for separate univariate meta-analyses. Moreover, the interpretation of the parameters is not straightforward. The variance at the study level does not refer to the total between-study variance for a given outcome (as in a univariate or multivariate meta-analysis). Rather, it refers to the between-study variance in the mean effect over outcomes. The less outcomes covary, the larger the between-study variance for given outcomes will result in differences between multiple effect sizes within studies, so the larger the variance will be at the outcome level. We have shown that the between-study variance in the study mean effects is equal to the covariance between effects observed within the same study.

Finally, the three-level model makes some assumptions that might be difficult to assess. The model assumes that outcome and study effects can be regarded as a random sample from a population of effects, or from a Bayesian perspective that effects are exchangeable, possibly conditional on the effects of covariates (Higgins, Thompson, & Spiegelhalter, 2009; Raudenbush, 2009). This assumption might be violated for instance due to reporting or publication bias. In our example, we found a negative correlation over studies between the number of outcomes reported and the mean effect, suggesting the existence of reporting bias. Furthermore, a normal distribution is assumed for the residuals at each of the levels. Raudenbush and Bryk (2002) show how the normality assumptions about these population distributions could be checked. If the normality assumption cannot be made, nonparametric estimation procedures could be used, such as the error bootstrap in which bootstrap samples are obtained by sampling from the estimated residuals at each of the levels, or the cases bootstrap in which studies and outcomes within studies are sampled together with the corresponding values for the dependent and independent variables (see Van den Noortgate & Onghena, 2005, for a description of these and other bootstrap procedures for meta-analysis). Another assumption is that the outcomes have a common between-study variance and that the between-outcome covariance is the same for each pair of outcomes. These are strong assumptions, but our simulation study suggests that the approach is relatively robust for a violation of these assumptions. Moreover, as mentioned above, it is possible to define different variance parameters at the second and third level of the three-level model for different kinds of outcomes, but this approach has not been studied in this study.

An important conclusion of the simulation study is that the traditional random effects meta-analytic model, which is equivalent to a two-level model, performs poorly unless the multiple effect sizes from the same study are truly independent. By ignoring the dependence in effect sizes, the two-level model can result in too small standard errors, and therefore in (largely) deflated coverage proportions of the confidence intervals. We conclude that if there are multiple outcomes per study, it is important to account for a possible dependence in the effect sizes.

Results of our simulation study further show that the three-level model performs as hoped for: standard errors for the mean effect estimate are relatively accurate, as well as the coverage proportions of the 90 % and 95 % confidence intervals. Regarding the variance estimates, we found that the total variance in effect sizes is divided over the outcome and the study level, according to the principle that the variance at the study level reflects the covariation between multiple outcomes from the same study. We see, however, that for the analysis of standardized mean differences, the total variance is somewhat underestimated. This negative bias can be explained by the dependence of the weights on the (observed) effect size, resulting in a downweighting of large effect sizes. This negative bias was also observed by Van den Noortgate and Onghena (2005), who described parametric and nonparametric bootstrap procedures that can be used to correct for this and other biases in parameter estimates. This bias depends on the effect size metric that is used. For instance, the negative bias was not observed by Van den Noortgate and Onghena (2003), who simulated theoretical effect size values directly from a normal sampling distribution with a variance that does not depend on the size of the effect.

These patterns are observed for each of the situations discussed in the paper. More specifically, we found similar results if the number of outcomes vary over studies, if the between-study variance depends on the outcome, if correlations vary over pairs of outcomes, if each study measures an own set of outcomes randomly drawn from a population of outcomes, or if all these five extensions are combined. Standard errors are quite accurate in all conditions, as well as the coverage proportions of the confidence intervals. Interestingly, whereas we found in the reference conditions that the standard errors for the mean effect are slightly overestimated when the effect varies over outcomes and outcomes are independent, this small positive bias disappears when not a common set of outcomes is measured in each study, but rather each study uses its own set of outcomes from a population of outcomes. The meta-analysis of Geeraert et al. (2004) about early prevention programs for child abuse and neglect resembles this situation: the authors found that there

was no consistency in the outcome variables used in the primary studies because child abuse and neglect is very difficult to observe directly. In other domains, there might be more consistency in the outcomes used. For instance, Rosa-Alcázar, Sánchez-Meca, Gómez-Conesa, and Marín-Martínez (2008), investigating the effect of psychological treatment of obsessive compulsive disorder, found that primary studies often looked at the same four types of outcome: obsessions and compulsions, general anxiety, depression and social adjustment, although they found that some studies also reported results for one or more other outcome variables. Whereas our simulation results suggest that standard errors might be slightly conservative if we have a common set of outcomes, and highly accurate if each study uses its own set of outcomes, real meta-analyses are typically situated in between.

Several limitations of the simulation study have to be mentioned. A first limitation is that results in principle are restricted to the conditions of the simulation design. For instance, the performance of the three-level might be different for smaller or larger data sets than the ones we simulated, or for other parameter values. Although we found similar results if we consider two groups of outcomes with high covariation within groups and small covariance between groups instead of a common covariance between each pair of outcomes, this does not imply that the results would be similar for other kinds of covariance structures as well. Our simulation study further focused on standardized mean differences. Findings for other kinds of effect sizes might be somewhat different; we expect for instance that the negatively biased total variance estimates is to a large extent due to using a standardized mean difference, as explained before. Still, we tried to simulate data for a variety of situations, and found that the patterns were similar in each of these situations, suggesting that results might be generalized to a certain extent. Moreover, results are in line with the results of another simulation study of Van den Noortgate et al. (2013), who investigated many more conditions. Another limitation of the simulation study is that data were simulated from normal distributions, therefore not violating the normality assumption that is made when using maximum likelihood estimation procedures. Finally, in our simulation study, we focused on an empty model, this is a model without predictor variables.

We conclude that if there are multiple outcomes per study, it is important to account for a possible dependence in the effect sizes. The three-level approach has been proven to be an appealing and reliable approach in a variety of situations, although further research is needed to assess the robustness of the approach in other situations, for instance in situations where in some studies multiple outcomes are measured in a single sample whereas in other studies outcomes are measured in independent samples, where the correlations between outcomes vary over studies, or where data are highly unbalanced, for instance when the large majority of studies report only one effect size whereas a few studies report many effect sizes. We are currently executing additional simulation studies to compare the performance of the three-level approach and the RVE approach, also for testing the effect of moderator variables.

**Acknowledgments** The simulation study was performed on the High Performance Cluster of the Flemish Supercomputer Centre.

## Appendix A: Proof that the variance at the study level is the covariance

Suppose that two observed effect sizes, $d_{jk}$ and $d_{j'k}$, stem from the same study, study k. According to Equation 6, $d_{jk}=\gamma_{00}+u_{0k}+v_{jk}+r_{jk}$ and $d_{j'k}=\gamma_{00}+u_{0k}+v_{j'k}+r_{j'k}$

Therefore,

$$\sigma_{d_{jk}d_{j'k}} = \sigma_{\left(\gamma_{00}+u_{0k}+v_{jk}+r_{jk}\right)\left(\gamma_{00}+u_{0k}+v_{j'k}+r_{j'k}\right)}$$

Because $\gamma_{00}$ is a constant, and adding a constant to one or both random variables does not affect their covariance, this covariance equals:

$$\sigma_{d_{jk}d_{j'k}} = \sigma_{\left(u_{0k}+v_{jk}+r_{jk}\right)\left(u_{0k}+v_{j'k}+r_{j'k}\right)}$$

The covariance between two linear combinations is described by Mood, Graybill and Boes (1974, p. 179):

$$\operatorname{cov}\left[\sum_{1}^{n} a_i X_i, \sum_{1}^{m} b_j Y_j\right] = \sum_{1}^{n}\sum_{1}^{m} a_i b_j \operatorname{cov}\left[X_i, Y_j\right]$$

Hence:

$$\sigma_{d_{jk}d_{j'k}} = \sigma_{r_{jk}r_{j'k}} + \sigma_{r_{jk}v_{j'k}} + \sigma_{r_{jk}u_{0k}} + \sigma_{v_{jk}r_{j'k}} + \sigma_{v_{jk}u_{0k}} + \sigma_{v_{jk}v_{j'k}} + \sigma_{u_{0k}r_{j'k}} + \sigma_{u_{0k}v_{j'k}} + \sigma_{u_{0k}u_{0k}}$$

Because in a multilevel model two residuals at the same level are assumed to be independent, as are residuals at two different levels,

$$\sigma_{d_{jk}d_{j'k}} = \sigma_{u_{0k}u_{0k}} = \sigma_u^2$$

## Appendix B: SAS Codes for the Example

Data set format

For the multilevel analyses, the data set should contain one row for each observed effect size. For our example, we prepared such a data set, called ABUSE, with the following variables (the dataset is available upon request from the first author):

STUDY: a study indicator with values from 1 to 39,
OUTCOME: an outcome indicator with values from 1 to 587,
ES: the effect size expressed as bias corrected standardized mean differences,
W: the inverse of the estimated sampling variance for each observed effect size, and
X: an indicator variable for the two groups of outcomes (1 refers to the outcomes directly related to child abuse and neglect, 2 to outcomes related to risk factors).

The first ten rows are given below:

| STUDY | OUTCOME | ES | W | X |
|---|---|---|---|---|
| 1 | 1 | -0.57426 | 5.5113 | 1 |
| 1 | 2 | 0.00000 | 10.5497 | 1 |
| 1 | 3 | -0.20948 | 4.2576 | 1 |
| 1 | 4 | -0.08965 | 10.1033 | 1 |
| 1 | 5 | -0.30052 | 5.8668 | 1 |
| 1 | 6 | 1.14184 | 1.3885 | 1 |
| 1 | 7 | 1.44236 | 1.4494 | 1 |
| 1 | 8 | 0.00000 | 4.9823 | 2 |
| 2 | 9 | 0.69606 | 3.8113 | 1 |
| 2 | 10 | 0.18574 | 27.4392 | 1 |
| | | ... | | |

Two-level meta-analysis

For the random effects two-level analysis, the following code is run:

```
proc mixed data=ABUSE method=reml;
      class STUDY OUTCOME;
      model ES= /solution ddfm=satterthwaite;
      weight W;
      random intercept/sub=OUTCOME;
      parms 1 1 /hold= 2;
run;
```

The *Proc Mixed*-command calls the mixed procedure for multilevel or linear mixed models. The data set is defined, and we ask for using the restricted maximum likelihood (REML) estimation procedure.

The *Class*-statement is used to define the categorical variables of our model, in our case the study and outcome indicators. In the *Model*-statement we define the model: the dependent variable (ES) on the left side of the equality sign, the predictor or moderator variables on the right. An intercept is included by default. In this random effects model, there are no moderator variables. The *Solution*-option requests the parameter regression coefficient estimates and tests in the output. The *ddfm=Satterthwaite*-option performs a general Satterthwaite approximation for the denominator degrees of freedom for the tests of the regression coefficients.

We use W, the inverse of the sampling variance, to weight the observed effect sizes in the analysis. However, weights that are used in the multilevel analysis will not only be based on the sampling variance, but also will automatically account for the estimated population variance(s) defined further. More specifically, the weights are equal to the inverse of the sum of the different sources of variance. The *Random*-statement specifies that the intercept varies randomly over outcomes. In the *Parms*-statement, we give starting values for the population variance of this intercept as well as for the residual variance.

We use the *Hold*-option to fix the second parameter to the starting value of 1. In this way and by using the inverse of the sampling variance as weights, the level-one variance is automatically fixed at the sampling variances that we defined. Using a more realistic starting value for the first parameter (e.g., the estimate from a previous analysis) can speed up the estimation. Finally, we close the code using the *Run*-statement, and we submit the code.

For the mixed effects two-level analysis, the code is adapted as follows:

```
proc mixed data=ABUSE method=reml;
      class STUDY OUTCOME X;
      model ES= X /solution noint ddfm=satterthwaite;
      weight W;
      random intercept/sub=OUTCOME;
      parms 1 1 /hold= 2;
      estimate 'group' X 1 -1;
run;
```

First, the X-variable is defined as a categorical variable by means of the *Class*-statement. Second, in the *Model*-statement we include the X-variable as a predictor. To get an estimate of the mean effect for each level of the X-variable, we drop the intercept, by using *noint* as an option in the *Model*-statement. Finally, to estimate and test the difference between both groups of outcomes in the expected effect, we use the *Estimate*-statement. The difference-parameter is labeled 'group', and is defined by a contrast with weights 1 and -1.1

Three-level meta-analysis

Because in the three-level random effects model we assume that the intercept might not simply vary over the 587 outcomes, but that there might be systematic differences between studies due to covariation between effect sizes from the same study, we include a second *Random*-statement:

```
proc mixed data=ABUSE method=reml;
      class STUDY OUTCOME;
      model ES= /solution ddfm=satterthwaite;
      weight W;
      random intercept/sub=STUDY;
      random intercept/sub=OUTCOME;
      parms 1 1 1 /hold= 3;
run;
```

We now have three sources of variance: between studies, between outcomes within studies, and sampling variance. In the *Parms*-statement, we define starting values for the three variances, and constrain the last one to 1.

The code for the random effects model is extended to the code for a mixed effects model in much the same way as for the two-level models.

## References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research, 82,* 436–476.

Arends, L. R., Voko, Z., & Stijnen, T. (2003). Combining multiple outcome measures in a meta-analysis: An application. *Statistics in Medicine, 22,* 1335–1353.

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.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1,* 97–111.

Cheung, M.W.-L. (2013). Modelling dependent effect sizes with three-level meta-analyses: A structural equation modelling approach. *Psychological Methods.* Advance online publication.

Cheung, S.F., & Chan, D.K.-S. (2014). Meta-analyzing dependent correlations: AnSPSS macro and an R script. *Behavior Research, 46*, 331–345.

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

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.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York: 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., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation of meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1,* 39–65.

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A, 172,* 137–159.

Hox, J. (2002). *Multilevel analysis. Techniques and applications*. Mahwah, NJ: Erlbaum.

Ishak, K. J., Platt, R. W., Joseph, L., & Hanley, J. A. (2008). Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine, 27,* 670–686.

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine, 30,* 2481–2498.

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

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

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS® system for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1,* 85–91.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Osburn, H. G., & Callender, J. C. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology, 77,* 115–122.

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

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). London: Sage Publications.

Riley, R. D. (2009). Multivariate meta-analysis: The effect of ignoring within-study correlation. *Journal of the Royal Statistical Society, Series A, 172,* 789–811.

Rosa-Alcázar, A. I., Sánchez-Meca, J., Gómez-Conesa, A., & Marín-Martínez, F. (2008). Psychological treatment of obsessive-compulsive disorder: A meta-analysis. *Clinical Psychology Review, 28,* 1310–1325.

Scammacca, N., Roberts, G., & Stuebing, K.K. (2013). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research.* Advance online publication.

Stevens, J. R., & Taylor, A. M. (2009). Hierarchical dependence in meta-analysis. *Journal of Educational and Behavioral Statistics, 34,* 46–73.

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.

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

Van den Noortgate, W., & Onghena, P. (2005). Parametric and nonparametric bootstrap methods for meta-analysis. *Behavior Research Methods, 37,* 11–22.

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta regression. *Statistics in Medicine, 21,* 589–624.