Behav Res (2013) 45:576–594
DOI 10.3758/s13428-012-0261-6

Three-level meta-analysis of dependent effect sizes

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

Published online: 9 October 2012

Abstract Although dependence in effect sizes is ubiquitous, commonly used meta-analytic methods assume independent effect sizes. We describe and illustrate three-level extensions of a mixed effects meta-analytic model that accounts for various sources of dependence within and across studies, because multilevel extensions of meta-analytic models still are not well known. We also present a three-level model for the common case where, within studies, multiple effect sizes are calculated using the same sample. Whereas this approach is relatively simple and does not require imputing values for the unknown sampling covariances, it has hardly been used, and its performance has not been empirically investigated. Therefore, we set up a simulation study, showing that also in this situation, a three-level approach yields valid results: Estimates of the treatment effects and the corresponding standard errors are unbiased.

Keywords Multilevel · Meta-analysis · Multiple outcomes · Dependence

Three-level meta-analysis of dependent effect sizes

Meta-analysis refers to the statistical integration of the quantitative results of several studies. For instance, if several experimental studies examine the effect of student coaching on students' achievement, a meta-analysis can be used to investigate how large the effect is on average, whether this effect varies over studies, and whether we can find factors that are related to the size of the effect (for an introduction to meta-analysis, see Borenstein, Hedges, Higgins, & Rothstein, 2009). Most often, the effects that were observed in the primary studies are not exactly the same as the ones predicted on the basis of the meta-analytic model. Most meta-analytic methods assume that these deviations of the observed effect sizes from their expected values are independent. This means, it is assumed that one specific observed effect size does not give information about the direction or size of deviation of another observed effect from the value we would expect on the basis of the meta-analytic model. Nevertheless, in practice, there are numerous sources of effect size dependence, and therefore, meta-analysts are very often confronted with dependent effect sizes. For instance, if several studies were done by the same research group, it is not unlikely that the effect sizes from this research group are more similar than effect sizes from two different research groups, because the effect sizes might be influenced by common factors, including the way the dependent and independent variables were operationalized, the characteristics of the studied population, the way subjects were sampled, and the observers or interviewers who collected the data (Cooper, 2009). In the same way, study results from the same country might be more similar than study results from different countries, again inducing dependence in observed effects.

Besides dependence over studies, there also might be dependence within studies. For instance, a specific effect can be investigated for several samples in a single study. In this situation, factors such as the ones described above possibly vary less within the same study than over studies, possibly resulting in dependence.

The dependence within studies is most obvious if effect sizes are based on a common sample (Becker, 2000; Gleser & Olkin, 1994). Studies often do not use one single outcome

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

J. A. López-López · F. Marín-Martínez · J. Sánchez-Meca
Universidad de Murcia,
Murcia, Spain

variable but, instead, report the treatment effect on different variables. These can refer to different but related constructs (such as anxiety and depression), different aspects of a construct (such as subscales of a psychological test), different operationalizations of the same construct (such as two instruments for anxiety), or repeated measures of the same construct (e.g., a posttest and follow-up measures). Another example of overlap in the samples is the case where multiple (independent) treatment groups are compared with a common control group.

Stevens and Taylor (2009) discussed another source of dependence, one that is not intrinsic to the design or the data but is related to the way the effect sizes are calculated and occurs when mean differences from independent groups are standardized using a common within-group standard deviation estimate—more specifically, the root *MSE* resulting from an ANOVA for a between-subjects factor with more than two levels.

Dependent effects sizes are less informative than independent effect sizes. Suppose that two outcome variables are perfectly correlated. Essentially, this means that both outcomes refer to the same latent variable and that effect sizes calculated for both outcomes will give exactly the same information. If, in a meta-analysis, both of these effect sizes are included as independent effect sizes, the same information therefore is used twice. In general, when outcome variables are correlated, information regarding one outcome overlaps with information yielded by the other outcome. By "inflating" the available information in this way, we will overestimate confidence in the results of the meta-analysis. In statistical terms, standard errors of the parameter estimates are likely to be underestimated, resulting in too small confidence intervals and too high a number of incorrect rejections of the null hypotheses (Becker, 2000). There is also another problem when ordinary meta-analytic methods are used to combine dependent effect sizes: Studies or groups of studies with multiple effect sizes will have a larger influence on the results of the meta-analysis than will studies reporting only one effect size, potentially resulting in biased estimates. For both reasons, it is important that the dependence between effect sizes be taken into account in our meta-analysis.

Yet the dependence over studies and, especially, the dependence within studies including independent samples are often overlooked, although recently, Stevens and Taylor (2009) and Hedges, Tipton and Johnson (2010) discussed the use of random effects for dealing with this kind of dependence. Some authors even explicitly have stated that if effect sizes are based on independent samples, effect sizes can be regarded as independent even if they stem from the same study (e.g., Littell, Corcoran, & Pillai, 2008).

The dependence within studies due to overlapping samples has received much more attention. Becker (2000) and Littell et al. (2008) have given an overview of approaches to accounting for this kind of dependence, as well as of the corresponding problems or limitations. In general, approaches boil down to three types: ignoring dependence, avoiding dependence, and modeling dependence. Due to reasons described above, ignoring dependence in principle is not appropriate. Yet if only one or two studies in a large set of studies based more than one effect size on the same sample, treating the effect sizes as independent will probably not substantially influence the results of the meta-analysis. In any case, if a researcher decides to treat multiple effect sizes from the same studies as independent, sensitivity analyses are recommended, comparing the results of the analysis ignoring the dependence with one or more alternatives (Becker, 2000).

A second strategy is to avoid dependence, which means that we restrict our analysis to one effect size per study. If more or less the same outcome variables are measured in each study, separate meta-analyses can be performed for each type of outcome. Also, this approach is not without problems or disadvantages. For instance, Rosa-Alcázar, Sánchez-Meca, Gómez-Conesa and Marín-Martínez (2008), investigating the effect of psychological treatment of obsessive–compulsive disorder, performed five separate meta-analyses for five types of outcome variables. Yet they found that only 7 of the 24 studies reported the effect of psychological treatment on social adjustment, a number that is too small for accurately estimating the between-study variance. Performing separate analyses even can become unfeasible if there are a lot of different outcomes and, especially, if chosen outcomes vary a lot over studies. Moreover, when separate meta-analyses are performed for different outcomes, testing differences between outcomes in the mean treatment effect or in the moderating effects of study characteristics is not straightforward. Finally, each separate meta-analysis uses only a subset of the data, which in principle results in less accurate estimates and less power in statistical tests (Gleser & Olkin, 2009).

Another way of avoiding dependence is not taking the individual effect sizes as the *units of analysis* but, rather, the samples (i.e., using one effect size per sample), the studies, or even research groups (Cooper, 2009). To obtain one effect size per higher level unit, one of the observed effect sizes from that unit can be selected, either randomly or because the chosen effect size refers to the outcome variable that is most of interest from a substantive point of view. A common way to reduce multiple effect sizes per higher level unit (e.g., per study) is to average them, an approach that especially makes sense if outcomes refer to the same construct. However, by simply averaging effect sizes within studies, the variance between effect sizes is artificially reduced, and informative differences between outcomes get lost (Becker, 2000; Cheung & Chan, 2008).

The third and most complex strategy is to model the dependence. Raudenbush, Becker and Kalaian (1988) proposed a multivariate model for analyzing multivariate effect size data. This model was extended by Kalaian and Raudenbush (1996) to a multivariate mixed model, allowing one to model variation between and within studies. The model, which will be presented later, assumes that effect sizes within the same study are possibly correlated. Whereas an ordinary meta-analysis uses sampling variance estimates of the effect sizes to approximate the optimal weights for the effect sizes and to estimate unknown parameters and corresponding standard errors, a multivariate meta-analysis uses the estimated sampling covariance matrix of the multivariate effect sizes. The multivariate approach has several advantages: The treatment effect is estimated for each outcome, all available information is used in one single analysis, it is possible to test contrasts between treatment effects of outcomes (e.g., whether the effect on a first outcome is the same as the effect on two other outcomes), and, similarly, it is possible to test differential moderating effects of study characteristics. The approach, however, is more complex to use, because it is not always implemented in software for meta-analysis. A major disadvantage of the approach is that for estimating the covariance matrix of the effect sizes, the correlations between the outcomes are also needed. If the results of standardized tests are used, correlations reported in the test manuals might give an idea about these correlations, but otherwise estimates of these correlations are difficult to obtain, because only rarely are estimates reported by the primary studies and the raw data that could be used to estimate the correlations are typically not available to the meta-analyst. Therefore, multivariate meta-analyses are seldom used or are used only on a subset of outcomes for which relatively accurate intercorrelation estimates are available. In addition, the multivariate approach, as well as the approach of separate meta-analyses for each outcome variable, is feasible only if there are only a few different outcomes.

In this article, we describe and evaluate an alternative approach to accounting for dependence within and over studies: the use of multilevel models. First, we briefly introduce multilevel models and their application for meta-analysis. Next, we give some examples of meta-analyses using three-level models, accounting for different kinds of dependencies. In one of the examples, a three-level model is used to account for sampling covariation. Because applying the multilevel approach in this situation is not straightforward and has not been validated before, we explore its performance with an in-depth discussion of the analysis results for a simulated data set and by presenting the results of an extensive simulation study. The article closes with a discussion and conclusions.

Multilevel meta-analysis

In social and behavioral sciences, data often have a clustered structure. For instance, if in educational research, first a set of schools is sampled from a population of schools and, in a second stage, students are sampled from the selected schools, data are clustered: The students participating in the study can be grouped according to the schools they belong to, as illustrated in Fig. 1.

This structure can induce dependence in the data: Students from the same school are, in general, more alike than students from different schools—for instance, due to (self-) selection effects or to effects schools have on their students. Therefore, school membership has to be taken into account when statistical analyses that assume independent residuals are performed.

Multilevel models have been developed to deal with such grouped data (Goldstein, 1987; Raudenbush, 1988). A simple two-level model (with a within- and a between-group level) includes a regression equation (Eq. 1) regressing the dependent variable Y on a predictor X, describing the variation over units (referred to with index i) within groups (index j):

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}. \tag{1}$$

Regression coefficients also get an index j, meaning that they are allowed to vary over groups. This variation at the group level is described using additional regression equations, possibly including group characteristics as predictors. Equations 2 and 3 at this second level regress the within-group intercepts and slopes, respectively, on a group characteristic Z:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \tag{2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j}. \tag{3}$$

Residuals at each level are typically assumed identically and independently normally distributed with zero means, and being independent over levels, this is

$$e_{ij} \sim N\left(0, \sigma_e^2\right) \quad \text{and} \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \\ \sigma_{u_0u_1} & \sigma_{u_1}^2 \end{bmatrix}\right) \tag{4}$$

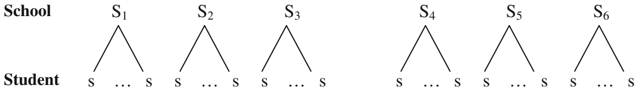

Fig. 1 A clustered structure in educational research

Especially when there is a large number of groups, the multilevel model is a very economic model, because regardless the number of groups that are included, the only parameters to be estimated are the regression coefficients at the group level (the γ's), and the (co)variances of the residuals at the first and second levels ($\sigma_e^2, \sigma_{u_0}^2, \sigma_{u_0u_1}, \sigma_{u_1}^2$). If one is interested in the individual level 1 regression coefficients (the β's), these coefficients can be estimated afterward using empirical Bayes techniques. By borrowing strength from the whole data set, empirical Bayes estimates are more efficient than the estimates based on the scores from the specific group alone (Raudenbush & Bryk, 2002).

An important strength of the multilevel model is its remarkable flexibility, allowing one, for instance, to define a third level describing the variation of the level 2 regression coefficients over groups of groups, including additional covariates at each of the levels, defining more complex covariance structures, such as an autocorrelation structure, or using a nonidentity link function for modeling discrete dependent variables, such as proportions or counts. Thanks to their flexibility, multilevel or mixed models have become popular in several fields, including education, psychology, economics, and biomedical sciences, for several kinds of nesting (students in schools, repeated measurement occasions within subjects, children in families, patients in doctors, etc.).

One application of multilevel models is meta-analysis (Hox, 2002; Raudenbush & Bryk, 2002). In a meta-analysis, we have indeed a similar nested data structure: Study subjects are nested within studies. Scores can vary over subjects from the same study (level 1), but there might also be differences between studies (level 2). For instance, if we want to combine the results of a set of treatment effectiveness studies comparing a control and a treatment group, we define a predictor at the subject level (X_{ij} of Eq. 1) as a dummy treatment indicator, having value 1 if subject i of study j belongs to the treatment group and 0 if the subject belongs to the control group. In this way, β_{0j} of Eq. 1 is equal to the expected value of a subject belonging to the control group in study j, whereas β_{1j} refers to the increase of the expected value if a study subject belongs to the treatment group and, therefore, can be interpreted as the treatment effect. One or more study characteristics can be included in the study-level regression equations (Eqs. 2 and 3) in an attempt to explain possible variation over studies in the expected baseline levels, β_{0j}, or in the treatment effects β_{1j}. If a study characteristic is found to explain or describe the treatment effects, the study characteristic has the role of a moderator variable.

There are, however, three major differences between an ordinary multilevel analysis and a typical meta-analysis. A first difference is that meta-analyses often combine results of studies in which the dependent variable is not always measured on the same scale, requiring a standardization of the data. To compensate for multiplicative factors, scores of each study can be divided by (an estimate of) the within-study standard deviation σ_{je}. More generally, linearly equatable scales can be made comparable by dividing the deviation of each score from the mean by the standard deviation. Note that whereas, for the unstandardized scores, β_{1j} is equal to the difference in expected values for both control and treatment conditions, by standardizing scores in either way,β_{1j} becomes the standardized mean difference that was proposed by Cohen (1988) and has become a popular effect size metric in social and behavioral research.

A second difference between ordinary multilevel analyses and meta-analyses is that a meta-analyst often does not have all raw data but, rather, depends on results reported in the form of summary statistics, test statistics, or effect size values. Fortunately, it is often possible to convert the results of each study to a common effect size metric and combine these effect sizes using a multilevel model. For instance, if, for each study, we can obtain a standardized mean difference, which is an estimate of the population standardized mean difference β_{1j}, Eq. 1 is adapted as follows:

$$\widehat{\beta}_{1j} = \beta_{1j} + r_j. \tag{5}$$

The residual r_j of Eq. 5 summarizes the effect of the residuals e_{ij} of all individual subjects in study j on the observed treatment effect. Differences over studies between treatment effects can be described in the same way as for raw data, regressing the β_{1j} on one or more study characteristics to investigate their moderating effects (Eq. 3). By substitution, we can write the two-level meta-analytic model (Eqs. 5 and 3) in one equation: $\widehat{\beta}_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j} + r_j$. If this second-level equation does not include any predictor, the multilevel meta-analytic model simplifies to the meta-analytic random effects model described by DerSimonian and Laird (1986) and Hedges and Olkin (1985), with $\widehat{\beta}_{1j}$ referring to an effect size estimate, γ_{10} to the overall effect, and $\sigma_{u_j}^2$to the systematic between-study variance. By including predictor variables, we assume that the variation of treatment effect over studies might be partly explained by moderating study characteristics, whereas the other part remains unexplained or random. Therefore, the multilevel meta-analytic model is also called a mixed meta-analytic model (Borenstein et al., 2009).

A third difference between raw data multilevel analyses and effect size meta-analyses is that for most commonly used effect size metrics, (a good approximation of) the sampling variance σ_{rj}^2can be calculated prior to performing the meta-analysis, and therefore, the variance of the residuals at the subject level does not have to be estimated in the meta-analysis anymore (Raudenbush & Bryk, 2002). We want to remark that the multilevel model (Eqs. 5 and 3) are applicable for other measures of effect size (such as log odds ratios or Fisher's *Z*-transformed correlation coefficients), as long as their sampling distributions are approximately normal, with a variance that can be estimated, and the same effect size metric is used for all studies.

Because a multilevel analysis of the raw data is similar to a two- or multistage analysis in which the subject-level regression coefficients (the βs from Eq. 1) are estimated and used as the dependent variable in new regression analyses, analyzing raw data or analyzing effect sizes (which can be regarded as standardized regression coefficients at the subject level) will give essentially the same results. An advantage of the analysis of the raw data (in medical sciences known as individual patient data), however, is that it allows for the incorporation of covariates at the subject level—that is, in Eq. 1—as discussed by Higgins, Whitehead, Turner, Omar, & Thompson, 2001. Moreover, using maximum likelihood estimation procedures, as is common for multilevel models, or using traditional estimation procedures for meta-analysis, such as the method of moments in DerSimonian and Laird (1986), will give very similar results, even if the normality assumption that is made for the maximum likelihood procedure is violated (López-lópez, Viechtbauer, Sánchez-Meca, & Marín-Martínez, 2010; Van den Noortgate & Onghena, 2003b). The major strength of using the multilevel modeling framework for meta-analysis is, however, its flexibility. In the following section, we will discuss one extension of the two-level meta-analytic model that is only very rarely used in practice: the inclusion of an additional level to account for dependence in the effect sizes.

Three-level meta-analyses for dependent effect sizes

Suppose that several teams performed more than one study, possibly resulting in dependent study results. The equation at the first level—that is, the subject level—states that due to sampling variation, the observed effect size from study *j* from team *k* possibly deviates from the "true" treatment effect for that study:

$$\widehat{\beta}_{1jk} = \beta_{1jk} + r_{jk} \qquad \text{with} \qquad \mathrm{r}_{jk} \sim N\left(0, \sigma^2_{r_{jk}}\right). \qquad (6)$$

Because the sampling variance, $\sigma^2_{r_{jk}}$, depends on the study—especially on the study size—it also gets an index for the study and the team. Additional equations state that the true treatment effect β_{1jk} can vary randomly over studies from the same team around a team-specific mean effect (θ_{10k} from Eq. 7, the second-level model) and that this team-specific mean effect, in turn, can vary randomly over teams around an overall mean effect (γ_{100} from Eq. 8, the third-level model):

$$\beta_{1jk} = \theta_{10k} + u_{1jk} \qquad \text{with} \qquad u_{1jk} \sim N\left(0, \sigma^2_u\right) \qquad (7)$$

$$\theta_{10k} = \gamma_{100} + v_{10k} \qquad \text{with} \qquad v_{10k} \sim N\left(0, \sigma^2_v\right). \qquad (8)$$

In order to try to explain this variation over studies and research teams, characteristics of these studies and teams can be included as predictors at the respective levels. It is even possible to allow the effect of a study characteristic to vary (partly) randomly over research teams.

The use of the three-level model is illustrated below using three examples of our own research. For all three analyses, use was made of the restricted maximum likelihood estimation procedure implemented in Proc Mixed from SAS (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Codes for running the analyses can be obtained from the authors.

Example 1: studies with multiple experiments

Van den Bussche, Van den Noortgate and Reynvoet (2009) performed a meta-analysis of the results of 54 studies, to assess the magnitude of psychological subliminal priming effects and to explore what factors are associated with the strength of the priming effects. In each study, subjects performed a number of tasks (e.g., indicating whether a number is larger or smaller than 5) after been given a congruent stimulus (e.g., a number larger than 5 is preceded by a number larger than 5) or an incongruent stimulus (e.g., a number larger than 5 is preceded by a number smaller than 5). Because the meta-analysis focused on subliminal priming, only studies were included that presented primes below the threshold for conscious perception—for instance, by presenting stimuli for very short durations. A priming effect is manifested as faster responses on congruent than on incongruent trials. For each subject, the priming effect can be estimated as the mean reaction time on incongruent trials minus the mean reaction time on congruent trials. The outcome used to quantify the observed priming effect in a primary study was defined as the mean difference over subjects, divided by the standard deviation of the differences.

In most of the 54 studies, more than one experiment was conducted, 156 in total, that differed from each other in one or more factors. Experiments from the same study cannot be regarded as independent, because they were done by the same researchers, and often with more similar research groups or more similar stimuli or procedures than in experiments from different studies. Therefore, we used a three-level model to analyze the data. We have a set of studies (level 3), experiments nested within studies (level two), and a sample of subjects for each experiment (level 1) (see Fig. 2). There are, therefore, three sources of variance: population differences between study population effects, population differences between effects of experiments from the same study, and, finally, sampling variance.

Two separate meta-analyses were performed on a set of 23 studies (88 experiments) containing semantic categorization tasks and a set of 32 studies (68 experiments) containing lexical decision and naming tasks. Using a three-level extension of the common meta-analytic random effects model—that is, a model without predictors at the experiment and study level—resulted for the first set of studies in a mean effect estimate of 0.80, but significant variance was found between studies, as well as between experiments within studies. Next, we investigated the moderating effect of several characteristics of the experiments and studies and found that, with only two predictors (prime novelty and category size), the variance at the experiment and study levels was reduced by 87 % and 40 %, respectively. For the second set of studies, a mean effect of 0.47 was found, and effects were found to vary over studies, but not over experiments, within studies. With two predictors (prime duration and target set size), the variance over and within studies was reduced by 99 % and 44 %, respectively.

Example 2: single-subject experimental designs

In single-subject experimental designs (SSEDs), subjects are measured repeatedly under different conditions—for instance, during a baseline phase without a treatment, a treatment phase, and a second baseline phase. The effect of the treatment for a subject can be evaluated by comparing the mean score of the subject during the baseline phase(s) and the mean during the treatment phase(s). To explore generalizability, SSED studies often include more than 1 subject, resulting in hierarchical data: Measurements are grouped or "nested" in subjects. If we have several SSED studies, subjects in turn are grouped in studies, meaning that three hierarchical levels can be distinguished (Fig. 3).

In SSED studies, data are typically reported using graphical displays, and conclusions are based on a visual analysis of these plots. The tradition of graphically presenting data means that the raw data can often be retrieved from the research reports, permitting one to combine the data of several subjects, using an ordinary two-level model to analyze data of one SSED study or an ordinary three-level model to combine the results of several studies (Van den Noortgate & Onghena, 2003a). This approach was illustrated by Van den Noortgate and Onghena (2008), reanalyzing the meta-analytic data set of Shogren, Fagella-Luby, Bae and Wehmeyer (2004), consisting of the results of 30 study subjects from 13 SSED studies that evaluated the impact of providing choice-making opportunities on problem behavior. In some of these studies, the intervention allowed subjects to choose the order of the task, whereas in other studies, subjects could make a choice between two tasks. *Problem behavior* refers to, for instance, aggressive behavior, destructive behavior, and off-task behavior. To compare results over studies, scores within studies were standardized by dividing the scores for each subject by the within condition standard deviation.

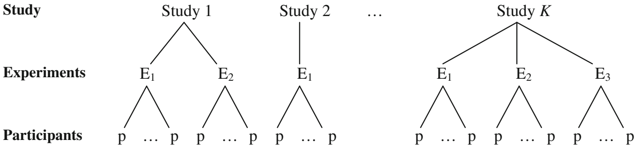

Fig. 2 Three-level hierarchical structure in the priming study meta-analysis

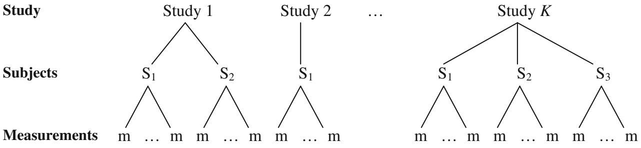

Fig. 3 Three-level hierarchical structure in a set of SSED studies

Results of the three-level meta-analysis revealed that, on average, the amount of problem behavior during the baseline phase was about 2.72 points, whereas the average standardized mean difference between phases was −1.72. This 63 % reduction of problem behavior was statistically significant, $z = 4.91$, $p < .0001$. Estimates of the between-study (co)variance of the baseline level and the choice effect were small and statistically not significant. Within studies, however, there was significant variation over subjects, both in the baseline level and in the choice effects. There was also some evidence for a negative covariation between baseline level and effect, suggesting that the intervention has a smaller effect for subjects with a relatively high baseline level of problem behavior, but this covariation was statistically not significant at the .05 significance level. Van den Noortgate and Onghena (2008) explored whether variation in baseline levels and in the effect sizes could be explained with a characteristic of the subjects (age) and a characteristic of the study (the kind of choice that was given), but neither of the two characteristics was found to have an effect and to reduce the variation.

Although this meta-analysis is atypical, in that raw data were analyzed rather than effect sizes, we found that combining effect sizes (more specifically, standardized regression coefficients for each subject) for this example gave almost identical parameter estimates and corresponding standard errors for the mean treatment effect, the variance of the treatment effect over subjects and studies, and the moderating effects of both predictors.

Example 3: multiple outcomes per sample

Geeraert, Van den Noortgate, Grietens and Onghena (2004) performed a meta-analysis on 40 studies evaluating the effect of early prevention programs for families with young children at risk for physical child abuse and neglect.

Whereas the ultimate goal of each of the evaluated early support programs was to reduce child abuse and neglect, the evaluation of a reduction of child abuse and neglect due to the program is not straightforward, because parents try to hide these aspects. We found that there were large differences between studies in the criteria used to evaluate the effect of the programs. Some studies used direct reports of child abuse or neglect by child protective services, some studies used 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. Most studies calculated the effect for more than one outcome variable. The number of effect sizes per study varied from 1 to 52, with an average of almost 15 effect sizes per study. A complicating factor for this analysis of dependent effect sizes, therefore, was that a lot of very different outcome variables were used in the studies. Even after categorizing the outcomes in broad categories, we ended up with 11 types of outcomes. Moreover, because the criteria do not refer to well-studied standardized measurement instruments, we did not have a clue about the correlation between these (types of) outcomes. In this situation, the use of a multivariate model is not feasible.

Therefore, Geeraert et al. (2004) used a three-level model, modeling the sampling variation for each effect size (level 1), variation over outcomes within a study (level 2), and variation over studies (level 3), as shown in Fig. 4.

Geeraert et al. (2004) found a small but statistically significant overall standardized mean difference, equal to 0.29, z = 6.59, p < .001, but also strong evidence for systematic differences between studies and within studies. Because a large part of the studies evaluated one specific program (the *Healthy Families America* program), an indicator was included to evaluate whether the effect for this intervention differed from the effect of other programs, but the difference was statistically not significant. A set of ten dummy variables was used to explore whether the type of outcome explained part of the variance within and between studies, but an omnibus test revealed that there were no differences between the 11 outcome categories.

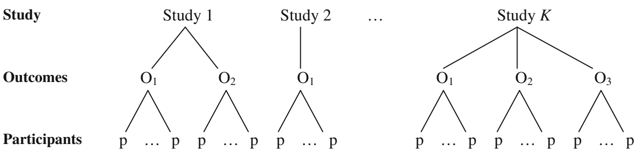

Fig. 4 Three-level hierarchical structure assumed for the child abuse meta-analysis

Whereas in the first two examples, the use of three-level models for meta-analysis is straightforward (for other examples with separate effect sizes for different subsamples, as in our first example, see Marsh, Bornmann, Mutz, Daniel, & O'Mara, 2009; Thompson, Turner, & Warn, 2001), this is not the case for the third example, in which several outcomes are based on the same sample. Indeed, it is clear from Fig. 4 that the structure assumed for the analysis does not correspond to the real data structure. Whereas, in reality, we have samples (level 1) nested in studies (level 2), with multiple effect sizes for each sample, the structure assumed in the three-level analysis implies that each sample corresponds to one outcome—in other words, that outcomes are calculated for independent samples. Therefore, the multilevel approach in principle incorrectly assumes that the resulting effect size measures are independent at the sample level. Yet it might be expected that the dependence between outcomes from the same study is taken into account by using an intermediate level of outcomes within studies. Although the multilevel approach is appealing, especially in situations in which we have a lot of outcome variables and/or correlations between outcomes are unknown, it has not been validated yet. In the remainder of the article, we will evaluate the three-level approach by comparing its results with those of a multivariate model and a traditional two-level model, by means of a simulation study.

An evaluation of three-level meta-analyses with sampling covariation

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful" (Box & Draper, 1987, p. 74). In this paragraph, we will evaluate the use of a meta-analytic model incorrectly assuming independent samples at the first level, but including an intermediate outcome-within-study level. To illustrate how parameter estimates are affected, we start with one large simulated data set and next look at the results of an extensive simulation study. All data were simulated using SAS and were analyzed using the restricted maximum likelihood procedure implemented in SAS Proc Mixed (Littell et al., 2006).

Although the focus of this article is on the meta-analysis of effect sizes, we do not simulate effect sizes directly but, rather, simulate raw data and calculate for each study a standardized mean difference. In this way, we can compare the results of an ordinary multilevel raw data analysis with those of a meta-analysis on effect sizes. Moreover, whereas the use of a multivariate model for effect sizes is difficult in practice because of a lack of information about the sampling covariation, a comparison of the results of a three-level analysis with those of a multivariate analysis on the raw data can give further insight into the three-level approach.

Understanding the three-level analysis

In order to understand better the interpretation of the results of a three-level analysis for effect sizes based on the same sample, one data set was simulated and analyzed, and results of the (two-level) multivariate model and of the (univariate) three-level model were compared. Data were simulated using the following bivariate model, with Y_{ijk} referring to the value for outcome j (j = 1, 2), for study subject i from study k:

$$\begin{cases} Y_{i1k} = \beta_{01k} + \beta_{11k}(Treatment)_{ik} + e_{i1k} \\ Y_{i2k} = \beta_{02k} + \beta_{12k}(Treatment)_{ik} + e_{i2k} \end{cases} \quad \text{with} \begin{bmatrix} e_{i1k} \\ e_{i2k} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2_{e_1} & \\ \sigma_{e_2e_1} & \sigma^2_{e_2} \end{bmatrix}\right) \tag{9}$$

$$\begin{cases} \beta_{01k} = \gamma_{010} + w_{01k} \quad \text{and} \quad \beta_{11k} = \gamma_{110} + w_{11k} \\ \beta_{02k} = \gamma_{020} + w_{02k} \quad \text{and} \quad \beta_{12k} = \gamma_{120} + w_{12k} \end{cases} \quad \text{with } \mathbf{w}_k \sim N(0, \Omega_w) \tag{10}$$

Because we have two outcome variables that possibly are related, the covariance matrix at the subject level (level 1) is a 2 × 2 matrix, with a positive value for $\sigma_{e_2e_1}$ indicating that if, in a study, a subject is sampled who has a high score on Y_1, this person is also likely to have a high score on Y_2. At the study level, we have a random residual for each outcome for the expected value in the control condition (w_{01k} and w_{02k}), as well as a random residual for each outcome for the treatment effect (w_{11k} and w_{12k}), and therefore, the covariance matrix at this level is a 4 × 4 matrix (Ω_w). The discussion and illustration of what happens if we use a three-level model ignoring sampling covariance, instead of a two-level multivariate model accounting for sampling variance, will be split up in three steps. First we discuss what happens to the parameter estimates if the mean effect is estimated, rather than the effect for each outcome variable separately. Second, we look at the effect of ignoring the sampling covariance (i.e., the covariance at the first level, the subject level), and finally, we discuss the use of a three-level model.

The parameter values that were used to simulate data for 300 studies (K = 300), each with two groups of size 50 (n = 50), are given in the second column of Table 1.

Model 1

The first model we used for analyzing the simulated data (model 1 from Table 1) is the multivariate model we used to generate the data (Eqs. 9 and 10). Parameter estimates are close to the parameter values used to generate the data.

Model 2

Sometimes, we might be interested not in the effect for each outcome separately but, rather, in the mean effect over outcomes. In the second model, we did not estimate the mean intercept and the mean effect for each outcome variable separately but, instead, estimated one overall intercept (γ_{00}, with $\beta_{01k} = \gamma_{00} + w_{01k}$ *and* $\beta_{02k} = \gamma_{00} + w_{02k}$) and one overall effect (γ_{10}, with $\beta_{11k} = \gamma_{10} + w_{11k}$ and $\beta_{12k} = \gamma_{10} + w_{12k}$). As a result (compare models 1 and 2), the between-study variance of treatment effects increases (because, in general, deviations of the treatment effects from the overall treatment effect become larger), but the between-study covariance decreases (because deviations within studies are less similar than before; one deviation increases, the other decreases). The estimated treatment effect now is in the middle of the previously estimated treatment effects for both outcomes.

Model 3

In the third model, we assume $\begin{bmatrix} e_{i1k} \\ e_{i2k} \end{bmatrix} \sim N\left(0, \sigma^2_e\right)$. We thus estimated only one variance parameter for both outcomes at the subject level (which is not a real restriction because, if the within-study variance cannot be assumed the same, scores are typically standardized). But more important, we assumed that the covariance is zero, whereas in fact, in the simulated data set, it is 0.79. Table 1 shows that as a result, the study level covariances between intercepts and between treatment effects become larger. This can be understood as follows. In the multivariate model, the intercept β_{0jk} refers to the expected value (or population mean) for outcome j in the control condition of study k, and this expected

Table 1 Parameter values and estimates (and standard errors) for a two-level multivariate data set

	Parameter values	Model 1 (multivariate)	Model 2 (no outcome indicator)	Model 3 (no level 1 covariance)
Fixed effects				
Intercept			−0.019 (0.020)	−0.019 (0.020)
Outcome 1	0.000	0.002 (0.022)		
Outcome 2	0.000	−0.009 (0.026)		
Treatment effect			0.209 (0.021)	0.209 (0.021)
Outcome 1	0.100	0.109 (0.025)		
Outcome 2	0.300	0.310 (0.025)		
(Co)variances				
Level 2 (study)				
Intercepts	$\begin{bmatrix}0.100 & \\ 0.050 & 0.100\end{bmatrix}$	$\begin{bmatrix}0.080 & \\ 0.040 & 0.112\end{bmatrix}$	$\begin{bmatrix}0.080 & \\ 0.040 & 0.111\end{bmatrix}$	$\begin{bmatrix}0.080 & \\ 0.055 & 0.111\end{bmatrix}$
Treatment effect	$\begin{bmatrix}0.100 & \\ 0.020 & 0.100\end{bmatrix}$	$\begin{bmatrix}0.089 & \\ 0.016 & 0.090\end{bmatrix}$	$\begin{bmatrix}0.098 & \\ 0.006 & 0.100\end{bmatrix}$	$\begin{bmatrix}0.098 & \\ 0.038 & 0.100\end{bmatrix}$
Level 1 (subjects)	$\begin{bmatrix}1.000 & \\ 0.800 & 1.000\end{bmatrix}$	$\begin{bmatrix}0.990 & \\ 0.790 & 0.994\end{bmatrix}$	$\begin{bmatrix}0.990 & \\ 0.790 & 0.994\end{bmatrix}$	0.992

Note. Standard errors of the estimates are given between parentheses. To simplify the table, at the study level, only the covariance between intercepts and the covariance between treatment effects is given, not the covariances between intercepts and treatment effects. Population values for these covariances were −0.025.

value varies over studies. Due to the fact that within a study only a sample of subjects is used, rather than the whole population, the observed mean is not necessarily the same as the expected value. More specifically, the sampling variation of the observed mean in the control condition for an outcome j is equal to $\frac{\sigma^2_{e_j}}{n}$. Because the residuals at both levels are independent, the total variance in the observed effect sizes is equal to the sum of the variance over studies in the expected control group level, and the sampling variance is $\sigma^2_{w_{0j}} + \frac{\sigma^2_{e_j}}{n}$. Similarly, the sampling covariance between the means for outcomes is equal to$\frac{\sigma_{e_je_{j'}}}{n}$, and the total covariance is $\sigma_{w_{0j}w_{0j'}} + \frac{\sigma_{e_je_{j'}}}{n}$. When the full multivariate model 1 is used, the observed variation and covariation in the intercepts is split up over the two levels. The results of the analysis show that if the covariance at the first level is not explicitly included in the model (model 3), the estimated sampling covariance in the control condition study means is added to the estimated covariance at the between-study level: $0.055 = 0.040 + \frac{0.790}{50}$.

The covariance in the treatment effects is affected in a similar way. The effect of the treatment dummy variable, β_{1jk}, refers to the difference in the expected values for the experimental and the control groups for outcome j in study k. The observed mean difference again varies, due to sampling variation and due to between-study variation. More specifically

$$\begin{aligned}&\overline{Y}_{.jk|experimental} - \overline{Y}_{.jk|control}\\&\quad= \left(\beta_{0jk} + \beta_{1jk} + \overline{e}_{.jk|experimental}\right) - \left(\beta_{0jk} + \overline{e}_{.jk|control}\right)\\&\quad= \gamma_{1j0} + w_{1jk} + \overline{e}_{.jk|experimental} - \overline{e}_{.jk|control}\end{aligned} \tag{11}$$

where $\overline{Y}_{.jk|experimental}$ and $\overline{Y}_{.jk|control}$ refer to the observed means in study k for outcome j in both conditions and $\overline{e}_{.jk|experimental}$ and $\overline{e}_{.jk|control}$ refer to the mean residual in both conditions.

Because residuals from the subject and study levels are independent and the residuals in both groups are independent, it is easy to show that the covariation between the mean differences for both outcomes is equal to $\sigma_{w_{1j}w_{1j'}} + \frac{2\sigma_{e_je_{j'}}}{n}$. Again, results of the example show that ignoring the sampling covariance (the covariance at the subject level) has as consequence that the estimated sampling covariation in the coefficients is added to the estimated between-study covariance: $0.038 = 0.006 + \frac{2{*}0.790}{50}$.

We conclude that, by ignoring the sampling covariation at the subject level, the covariation is still accounted for by overestimating the study level covariation by the same amount. Therefore, all other parameters and standard errors remain unchanged.

Model 4

In a third step, we look at the results of the three-level analysis, with samples, outcomes, and studies as the units at the respective levels. More specifically, we analyze the data using the following model:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}(Treatment)_{ik} + e_{ijk} \quad \text{with} \quad e_{ijk} \sim N\left(0, \sigma_e^2\right)$$

$$\begin{cases} \beta_{0jk} = \theta_{0k} + v_{0jk} & \text{with} \quad \begin{bmatrix} v_{0jk} \\ v_{1jk} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2_{v_0} & \\ \sigma_{v_0v_1} & \sigma^2_{v_1} \end{bmatrix}\right) \\ \beta_{1jk} = \theta_{1k} + v_{1jk} & \end{cases}$$
$$\begin{cases} \theta_{0k} = \gamma_{00} + u_{0k} & \text{with} \quad \begin{bmatrix} u_{0k} \\ u_{1k} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2_{u_0} & \\ \sigma_{u_0u_1} & \sigma^2_{u_1} \end{bmatrix}\right) \\ \theta_{1k} = \gamma_{10} + u_{1k} & \end{cases} \tag{12}$$

Equation 12 states that the intercept and treatment effect of outcome *j* and study *k* can (co)vary over outcomes within studies, as well as over studies. Whereas β_{0jk} and β_{1jk} refer to the control group level and treatment effect for outcome *j* in study *k*, θ_{0k} and θ_{1k} refer to the mean control group level and treatment effect (over outcomes) in study *k*, and γ_{00} and γ_{10} to the overall control group level and effect (over outcomes and studies). Parameter estimates for Eq. 12 are given in Table 2 (model 4).

A comparison of the results of the three-level model (model 4 from Table 2) and the multivariate model without a covariance at the first level (model 3 from Table 1) reveals that in the three-level analysis, the variance in the intercepts from the multivariate model is split up in variance over studies and variance over outcomes from the same study: 0.055 + 0.040 = 0.095, which is in the middle of the between-study variances of the intercepts for both outcomes in the multivariate model (0.080 and 0.111). Similarly, the variance over studies of both treatment effects (0.098 and 0.100) from the multivariate model is split up in the three-level model into variance over studies and variance over outcomes: 0.038 + 0.061 = 0.099.

Table 2 Parameter estimates (and standard errors) using three-level models

	Model 4 (three-level model for raw data)	Model 5 (three-level model for effect sizes)
Fixed effects		
Intercept	0.003 (0.021)	
Treatment effect	0.210 (0.021)	0.207 (0.021)
(Co)variances		
Level 3 (study)	$\begin{bmatrix} 0.055 & \\ -0.035 & 0.038 \end{bmatrix}$	0.035
Level 2 (outcomes within studies)	$\begin{bmatrix} 0.040 & \\ 0.012 & 0.061 \end{bmatrix}$	0.061
Level 1 (subjects)	0.992	*

* For the effect size analysis, the sampling variance, $\sigma^2_{r_{jk}}$, depends on the study and is estimated before the actual meta-analysis is performed.

The part going to the between-study level is equal to the covariance between the outcomes: 0.055 for both intercepts, 0.038 for the treatment effects. This is not surprising: In a multilevel analysis with units within groups, the variance between groups is equal to the covariation within groups (Snijders & Bosker, 1999). Also, intuitively, this makes sense: The larger the covariation between outcomes, the smaller the differences are within studies, and the larger the differences are between study means. If outcomes do vary over studies but do not covary, we expect to find differences between outcomes in a study, but on average, there will be no difference between studies.

Importantly, the mean effect estimate and corresponding standard error from the three-level model (model 4) are exactly the same as those of the multivariate model (model 3, but also model 2).

Model 5

Finally, we summarized, for each outcome and study, the data in a standardized mean difference, estimated the corresponding sampling variance (Hedges & Olkin, 1985), and analyzed these effect sizes using a three-level meta-analytic model (Eqs. 6, 7, 8). Comparing the results (see Table 2, model 5) with those of model 4 illustrates that the meta-analysis of the effect sizes gives almost identical parameter estimates as the analysis of the raw data.

Simulation study

In order to evaluate in a more systematic way the three-level modeling approach, we simulated a number of data sets in the same way as the example data set. The following parameters were varied: the number of outcomes (J = 2 or 5), the covariance between outcomes at the subject level ($\sigma_{e_je_{j'}} = 0,\ 0.4,\ 0.8$), the covariance in treatment effects between outcomes at the study level ($\sigma_{w_{1j}w_{1j'}}$ = 0, 0.02 or 0.04), the overall mean effect size (0, 0.20, 0.40), the deviation of the outcome effects from the overall mean effect (all two or five outcomes have same effect vs. deviations of −0.20 and +0.20 for outcomes 1 and 2 or −0.20, −0.10, 0, 0.10, and 0.20 for the five outcomes), the number of studies (K = 30 or 60), and the group sizes (n = 25 or 50). Values for the mean effects were chosen to be representative for the small and moderate effects commonly found in social and behavioral sciences (Cohen, 1988). The value used for the between-study variance in effect sizes (0.10) results in a realistic ratio with the sampling variance of observed effect sizes (about 0.08 for studies with n = 25 and 0.04 for studies with n = 50) and avoids the possibility that the effects of correlation at either level become ignorable (Riley, 2009). Covariances at the subject level were chosen to cover the range of possible values of correlation coefficients (covariances correspond with correlation coefficients of 0, .4, and .8). Whereas a high correlation at the first level is not unlikely in reality (e.g., when outcomes may refer to different instruments for the same construct), correlation coefficients at the second level are likely to be small (Lipsey & Wilson, 2001). Therefore, we used relatively small covariances at this level, corresponding to correlation coefficients of 0, .20, and .40. Differences between the effects of the outcomes were chosen to be relatively large (e.g., for two outcomes and an overall mean of .20, the outcome-specific mean effects are equal to 0 and .40), to be sure that a possible effect on the performance of the different models would be visible.

In total, we therefore have 2 × 3 × 3 × 3 × 2 × 2 × 2 = 432 combinations. For each combination, we simulated 1,000 data sets, 432,000 in total. Each data set was analyzed using the multivariate model, but including only one parameter for the treatment effects for all outcomes (model 2 from Table 1), with a three-level model (model 4 from Table 2), and with a traditional meta-analytic model that ignored the dependence of outcomes within studies (Eqs. 5 and 3)—that is, a two-level model without the third level, the study level. In addition, we summarized the data for each outcome within each study by calculating standardized mean differences, using the simulated raw data, and analyzed the resulting effect sizes with three-level and two-level models equivalent to the raw data models.

We want to stress that if we have the raw data, the multivariate model can be used because the sampling covariance can be estimated using the raw scores, and there is no reason to use the three-level approach unless the number of outcomes is very large and we are simply interested in the mean effect. Yet, if we only have effect sizes, often not enough qualitative information is available regarding the sampling covariances, and a multivariate model is not applicable.

Results of the simulation study

A first important result is that for all conditions, parameter estimates regarding the treatment effects (mean effects and standard errors, as well as variance components) are almost identical when performing multilevel meta-analyses of the effect sizes, rather than performing multilevel analyses on the raw data. Therefore, only the raw data results will be discussed, but conclusions equally apply to the meta-analyses of effect sizes. Moreover, because we are interested in the treatment effects—more specifically, the average effect and the variation in treatment effects—we will not discuss the results for the intercepts (referring to the level of performance in the control conditions).

Second, regarding the mean effects, we did not find important bias in any of the conditions (with absolute bias estimates never exceeding 0.006), and therefore, bias estimates are not further discussed here. We also found that the estimates of the mean effects are very similar for each of the models, with correlations of .99 and more. Estimates of the overall effects were even exactly the same for the univariate two- and three-level models. As was expected, the *MSE* of the estimates decreases with an increasing number of studies, an increase in the size of studies, and a decreasing covariance at the subject or study level. Bias and *MSE* estimates can be found in the Appendix.

Of major interest in our simulation study are the estimated standard errors for the mean effects, because these are used to test the mean effects and to construct confidence intervals. We evaluated the standard error estimates in two ways. First, standard errors corresponding to a parameter estimate refer to the standard deviation of the sampling distribution of the parameter estimator. Because we simulated a large number of data sets for each condition, it can be assumed that the standard deviation of the parameter estimates within each condition is a good approximation of the standard deviation of the sampling distribution and, therefore, can be used as a criterion to evaluate the standard error estimates. Figure 5 shows the mean standard error estimates and the standard deviations of the estimates for n = 25, K = 50, a study level covariance of .02, a mean effect of 0.4, and two outcome variables (trends are similar for other conditions). When we have a common outcome effect (Fig. 5a), the standard deviation of the estimates is slightly larger when the multivariate model is used, as compared with using the multilevel model. This is because, in the multivariate model, more parameters are estimated using the same amount of data and, therefore, the estimates are more fluctuating. Nevertheless, the mean standard error estimates are slightly smaller for the multivariate model than for the three-level model. In sum, whereas the standard errors for the three-level model are accurate, for the multivariate model they are somewhat too small. It is also clear from the graph that the variation in parameter estimates increases with an increasing covariance: If outcomes covary

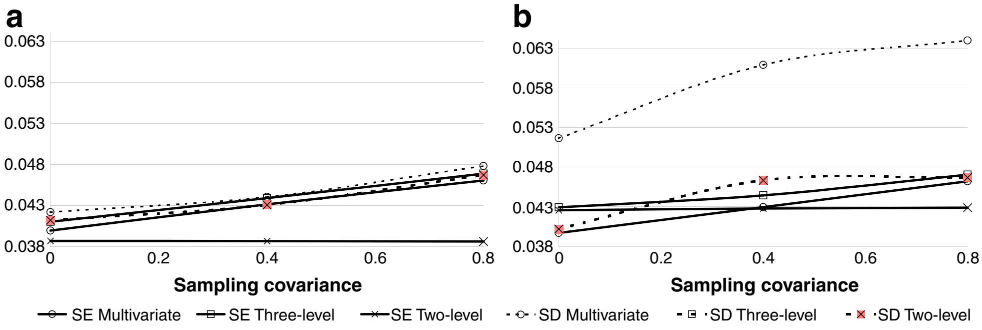

Fig. 5 Standard deviation (*SD*) of the overall effect estimates and mean standard error estimates (*SE*s) for the multivariate model, the three-level model, and the two-level model ignoring dependence. Left graph: same effect for all outcomes; right graph: outcome-specific effects

substantially, a large deviation of the scores on one outcome in a study is less likely to be compensated for by the scores on the other outcome. Standard errors for both the multivariate approach and the three-level approach follow this increasing trend. This is not the case if a two-level model is used: Standard errors do not depend on the sampling covariation. Therefore, whereas standard errors for this model are only slightly too small if there is no sampling covariation (due to between-study covariation of the outcomes), their negative bias become much more pronounced with increasing sampling covariance.

If the effect is not the same for each outcome but the same models are used to estimate the mean effect (Fig. 5b), estimates for the multivariate model vary much more. Yet standard errors of the multivariate and three-level models remain the same. Whereas standard errors of the three-level model are still relatively accurate, with only a small positive bias when there is no sampling covariance and a small negative bias if there is a strong sampling covariance, for the multivariate model they are severely negatively biased. The results suggest that the multivariate model does not work well for estimating the mean effect. The standard errors for the two-level model are larger than when there is no variation between outcomes (and can even be positively biased if there is no sampling covariation), but again they do not increase with increasing sampling covariance, resulting in negative bias if there is large sampling covariance.

A second way to evaluate the standard error estimates is by looking at the coverage proportion of the confidence intervals calculated using these standard error estimates. Table 3 gives the coverage proportions for 90 % confidence intervals for two and for five outcomes.[1] A dimension that is not included in the tables (but does not affect the coverage proportion) is the mean effect: In each condition, the coverage proportion is calculated over the three parameter values of the mean effect.

In line with the evaluation of the standard errors by comparing them with the standard deviation of the parameter estimates, we see that the coverage proportions for the 90 % confidence intervals are accurate for the three-level models in almost all combinations. Only if there is no covariation at either level and there are differences between outcomes is the coverage proportion slightly too high.

For the multivariate model, coverage proportions are, in general, slightly too small if the mean effect is the same for all outcomes, again suggesting that the standard errors are too small. When there are real differences between outcomes in the treatment effects (right side of the table), the coverage proportions are much too small for all combinations, suggesting that the multivariate model does not work well for estimating the mean effect. Also, for the two-level model, coverage proportions often are too small, unless outcomes do not covary at either level. As could be expected, conclusions are similar but more pronounced if we have five outcomes instead of two outcomes per study, resulting in much too small coverage proportions for the traditional two-level model.

If the mean effect is zero, the coverage proportion also refers to the proportion of confidence intervals including zero and, therefore, to the probability of not making a type I error. If the dependence is ignored, for some conditions, the proportion of type I errors is very substantially inflated (up to 36 %, instead of the nominal 10 % significance level; i.e., when there is substantial covariation at each level and we have five outcomes with a common mean effect).

The models give not only mean effect and corresponding standard error estimates, but also estimates of the between-study (and for the three-level model, the between-outcome) variance. In our simulation study, we saw the same patterns as in the elaborated simulated example discussed above: In the three-level model, the total variance between effect sizes

[1] For the data with five outcomes, we estimated only the meta-analytic models for combining effect sizes because raw data analyses, and especially the multivariate analysis, were very time consuming. Therefore, Table 3 does not include results for the multivariate model for five outcomes.

Table 3 Coverage proportions of the 90 % confidence intervals for the multivariate and the univariate three- and two-level models

			Common effect												Outcome-specific effect											
		$\sigma_{w_{1j}w_{1j'}}$	0				0.02				0.04				0				0.02				0.04			
		K	30		60		30		60		30		60		30		60		30		60		30		60	
J	$\sigma_{e_je_{j'}}$	n	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50
2	0	Multiv.	.87	.87	.88	.87	.86	.87	.88	.88	.86	.85	.89	.89	.79	.76	.81	.78	.77	.75	.80	.78	.77	.74	.78	.75
		3-level	.91	.92	.91	.90	.90	.91	.90	.90	.90	.89	.91	.91	.94	.94	.93	.93	.93	.93	.92	.93	.92	.91	.91	.91
		2-level	.89	.91	.89	.89	.88	.88	.87	.88	.86	.84	.87	.86	.93	.93	.93	.93	.91	.92	.92	.92	.90	.89	.89	.90
	0.4	Multiv.	.88	.87	.89	.89	.87	.85	.88	.89	.86	.86	.89	.88	.78	.75	.81	.76	.77	.75	.77	.77	.73	.72	.76	.74
		3-level	.91	.91	.90	.91	.90	.89	.89	.91	.89	.89	.90	.89	.91	.93	.91	.92	.91	.92	.89	.92	.90	.89	.89	.91
		2-level	.88	.88	.86	.89	.85	.84	.85	.86	.83	.83	.84	.84	.90	.92	.91	.91	.89	.90	.88	.90	.87	.86	.86	.89
	0.8	Multiv.	.87	.86	.87	.88	.87	.86	.89	.88	.86	.86	.88	.87	.76	.75	.78	.77	.74	.71	.76	.76	.72	.68	.74	.71
		3-level	.89	.90	.89	.89	.90	.89	.90	.90	.89	.89	.89	.89	.90	.92	.90	.90	.90	.90	.90	.91	.90	.89	.90	.91
		2-level	.83	.85	.83	.86	.82	.83	.83	.84	.80	.80	.80	.81	.89	.91	.88	.90	.87	.88	.86	.90	.86	.87	.86	.87
5	0	3-level	.92	.91	.90	.92	.90	.89	.89	.91	.90	.89	.90	.90	.92	.93	.92	.92	.91	.90	.91	.91	.89	.89	.90	.90
		2-level	.90	.89	.89	.90	.82	.81	.82	.82	.78	.73	.77	.75	.91	.92	.91	.91	.86	.83	.85	.85	.79	.76	.78	.77
	0.4	3-level	.90	.89	.89	.89	.89	.89	.90	.89	.91	.89	.90	.90	.90	.90	.89	.90	.90	.89	.90	.89	.89	.90	.90	.90
		2-level	.79	.81	.77	.82	.74	.75	.74	.74	.70	.68	.69	.69	.81	.85	.82	.86	.76	.77	.77	.77	.71	.72	.73	.72
	0.8	3-level	.90	.89	.89	.89	.90	.89	.90	.90	.90	.90	.90	.90	.89	.89	.90	.89	.90	.88	.90	.89	.90	.89	.91	.91
		2-level	.70	.75	.71	.74	.68	.69	.67	.70	.63	.65	.65	.64	.73	.78	.73	.79	.68	.72	.69	.73	.65	.68	.67	.69

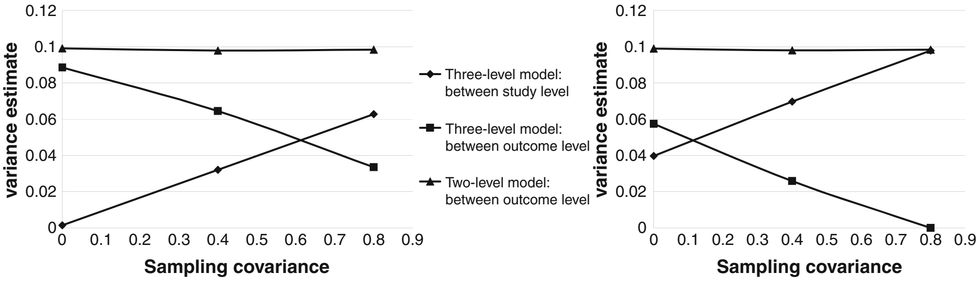

Fig. 6 Median variance estimates for the two- and three-level models if the covariance at the study level is zero (left) or 0.04 (right) (n = 25)

is split up into two parts, one referring to the variation between outcomes within a study, one referring to the variation between studies. Figure 6 shows this split-up; in the case in which the sample size n = 25, the number of outcomes is two, the treatment effect is common to all outcomes, and the study-level covariance between outcomes is equal to 0 (left graph) and 0.04 (right graph). As was discussed above, the variance between studies refers to the total covariation between outcomes, $\sigma_{u_{11}u_{12}} + \frac{2\sigma_{e_2e_1}}{n}$, which for the simulated conditions is equal to 0.00, 0.032, and 0.064 (left graph) and 0.040, 0.072, and 0.104 (right graph), depending on the sampling covariance. Figure 6 shows that in the three-level model, the variance estimate at the study level is an unbiased estimate of the total covariance between outcomes. Only in the condition where the total covariance between outcomes is equal to 0.104 is the estimated study-level variance slightly smaller. More specifically, the study-level variance does not exceed the between-study variance of the model used to simulate the data (0.100). Furthermore, Fig. 6 also shows that in the three-level model, the variance between studies (0.100) is split up into a variance at the study level and a variance at the outcome level, although the sum of both variance estimates is typically slightly smaller. This negative bias is especially visible if there is no covariance: The total variance is now only 0.090, instead of 0.100. Figure 6 also shows that the estimate of the between-study variance when using the two-level model (ignoring the dependence) is always almost equal to the between-study variance of the multivariate model (0.10). Similar patterns are found for other conditions.

Conclusions

In this article, we described the use of three-level models for dealing with dependent outcomes in meta-analysis and illustrated this approach using several real data examples. Whereas most traditional approaches for combining dependent effect sizes involve choosing a proper *unit of analysis* (Cooper, 2009), multilevel models were exactly developed "for the analysis of data sets comprising several types of unit of analysis" (Snijders, 2003, p. 674). Three-level models are suitable when studies are clustered—for instance, in research groups or countries—or when, within studies, we have several effect sizes calculated on independent groups, two situations in which dependencies are typically ignored. Sometimes, it is even explicitly stated that as long as there is no overlap between samples, multiple estimates from the same studies are truly independent (Littell et al., 2008). Our simulation study nonetheless shows that ignoring covariance at the study level also might result in biased standard errors and confidence interval coverage proportions, even if samples are independent.

An important conclusion is that, although the multilevel model we proposed for dealing with multiple outcomes within the same study in principle assumes no sampling covariation (or independent samples), our simulation study suggests that using an intermediate level of outcomes within studies succeeds in accounting for the sampling covariance in an accurate way, yielding appropriate standard errors and interval estimates for the effects. This is true for both the three-level model for raw data and the three-level model for analyzing effect sizes, which gave very similar results. In this article, we mainly presented the results for the analysis of raw data, because in this way we could compare the results with those of a multivariate approach. Whereas a multivariate approach, in principle, is also possible for the analysis of effect sizes, estimating the sampling covariance is often not feasible without having the raw data, making the multivariate approach not always feasible in practice. The simulation study also showed that using a random effects meta-analytic model assuming independent effect sizes might result in flawed inferences. This is especially true if the outcome variables are more correlated and the number of outcomes per study is higher. For instance, coverage proportions of the 90 % confidence intervals for the mean effects were typically between .65 and .75 if the number of outcomes was five, with intercorrelations of .80. This result is not unexpected: In general, the underestimation of the standard errors when clustered data are treated as independent data depends on the intraclass correlation (*ICC*) and the cluster size. For a two-level structure, for instance, Kish (1965) showed that the *design effect*—that is, the efficiency loss when cluster sampling is used for estimating a population mean rather than using simple random sampling—is equal to $1+(n-1)ICC$, with n being the cluster size. If the size of the clusters does not vary too much over clusters, the design effect is approximated well by replacing n by the average cluster size.

Although already, at the early development of multilevel analysis theory, the link was made with meta-analysis (Raudenbush & Bryk, 1985), meta-analysts seem to have failed to take advantage of the power of multilevel models for meta-analysis. Three-level analyses, for instance, are still very uncommon. We see nevertheless several important advantages to using a multilevel approach.

First, the multilevel model is a very flexible model. Whereas in meta-analysis the mixed effects meta-analytic model with a random study effect is often regarded as the ultimate model because it can account for moderator variables without making the strong assumption that these moderator variables explain all systematic variation between studies, the multilevel model is even more general than this two-level model, allowing one to define additional levels. In this way, it is possible to account for several sources of dependence at the same time. In our simulation study, we showed that a three-level model can account for sampling covariance, but the model can easily be extended with an additional upper level to account for a possible nesting of studies. If we have multiple outcomes based on the same sample, a multilevel model without predictors can be used to estimate the mean effect over all samples. Yet we can also include a characteristic of the outcomes as a predictor variable to explore whether the effect depends on the type of outcome. We even can include an outcome indicator as a predictor (as a set of dummy variables) to estimate the mean effect for each outcome, as is done in a typical multivariate model. Moreover, if we are interested in the treatment effect of each specific outcome, we could also perform separate meta-analyses, although an advantage of performing a multilevel meta-analysis with an outcome indicator over separate meta-analyses is that we can perform an omnibus test of differential mean effects—that is, a test of the effect of the outcome indicator. Moreover, contrasts can be estimated and tested—for instance, for evaluating whether the treatment effect is the same for two specific outcomes or for evaluating whether the effect for outcome one differs from the effect of outcomes four and five, tests that are less straightforward when performing separate meta-analyses for each outcome. The model can also easily be extended by including other covariates at each of the levels, in an attempt to explain variation at that level. An interaction term (or contrasts) can be used to estimate and test differential moderating effects of a predictor for various outcomes or for various types of outcomes. An important property of the multilevel model is that it does not require that the number of outcomes be the same for each study: If one study reports the effect for one outcome and another study the effect for five outcomes, all effects are used in the analysis. We further want to note that although, in our examples and discussion, we used standardized mean differences as the effect size metric, the multilevel meta-analytic model is equally appropriate for combining other commonly used metrics, such as odds ratios or Pearson's correlation coefficients. The approach assumes, as do most meta-analytic approaches, only that the sampling distribution of the metric is normal and, therefore, that a normalizing transformation of the effect sizes, such as the logarithmic transformation of odds ratios or the Fisher's Z transformation of correlations, might be required.

A second strength of the multilevel approach is that it is a relatively simple and intuitive way to account for dependencies. This is partly because multilevel models are discussed in several excellent handbooks (e.g., Hox, 2002; Raudenbush & Bryk, 2002) and software is widely available to use them—more specifically, specialized software, such as MLwiN and HLM, or general (statistical) software, such as SAS. Although depending on the situation, other approaches may be very helpful, these other approaches sometimes oversimplify the problem of dependency (e.g., by simply ignoring the dependency), inefficiently use only part of the data for each analysis, or are complex to implement because of a lack of necessary information, as described above.

A third strength of the approach is that multilevel models automatically account for the hierarchical structure in the data. If, for instance, one study results in 20 effect size estimates, this study will not contribute 20 times as much to the estimation of the mean effect, as compared with a study reporting only 1 effect size. Rather, this study is regarded as only one study yielding information about one study-specific mean effect in the distribution of study mean effects. The exact weight of each study will depend on the dependence between effect sizes from the same study: The smaller this dependence, the less the weight given to each of the individual effect sizes depends on the number of effect sizes reported in the study.

Specifically for accounting for sampling covariance, we see three major advantages of the multilevel approach for analyzing multivariate effect size data, as compared with a truly multivariate meta-analytic model. First, the multilevel approach does not require that the sampling covariance of the effect size estimates is "known" in advance. The covariation rather is taken into account by using the between-study variance as a "stand in" for this covariance. This is an important advantage, because the problem that, often, no or little information about the covariances is available is exactly the reason why multivariate meta-analyses are only seldom used. A second advantage is that because the sampling covariance is not to be known in advance, the multilevel model is also more applicable for metrics for which the formula for calculating the sampling covariance has a very complex form, gives biased estimates, or is unknown (see Becker, 2000, for a discussion of the multivariate distribution of commonly used effect sizes).

Finally, in the example of Geeraert et al. (2004), we saw that especially for dependent variables that are difficult to operationalize, there can be much variation between studies in the outcome variables used. Whereas this makes the multivariate model practically infeasible, the three-level model assuming a distribution of outcomes within studies is still easy to use. We recognize, however, that although the possibility of making estimates of the mean effect over outcomes is attractive, using the type of outcome or an outcome indicator as a predictor variable is to be preferred, for both conceptual and statistical reasons: If the effect varies over outcomes, the mean effect is less informative because effects might obscure each other. In addition, if one type of outcome is more often reported in the set of studies, the effect of this outcome will have a stronger influence on the average effect, possibly inducing bias. An example is where studies are less likely to report effect sizes for outcomes that are less affected by the treatment, resulting in a positive reporting bias. Moreover, the simulated data illustrate that including a moderator variable reduces standard errors, as well as the bias in the standard errors.

Although results of the simulation study are very promising, we are aware of some limitations of the study. Conclusions are, in principle, restricted to the conditions for which data were generated. More specifically, we generated data only for a relatively large set of studies (K=30 or 60). It is known from the literature on multilevel analysis (e.g., Maas & Hox, 2005) or multilevel meta-analysis (e.g., Van den Noortgate & Onghena, 2003b; Viechtbauer, 2005) that when (restricted) maximum likelihood procedures are used, smaller numbers of units at the highest level (in this case, the study level) might result in underestimated standard errors and, therefore, inflated type I error rates for testing the regression coefficients (in this case, the overall effect size and/or the moderator effects), but especially in biased estimates of the variance at the highest level and of the corresponding standard error. Because, in multilevel literature, 30 units is often regarded as the smallest acceptable number of units at the highest level (e.g., Kreft & De Leeuw, 1998), we did not simulate data sets with less than 30 studies.

Furthermore, we simulated only balanced data sets, with the same sample size and the same number of outcomes for each study. In practice, it often occurs that part of the studies report only one effect size, whereas another part of the studies report two or more effect sizes. In line with the design effect described by Kish (1965), we expect that also in this situation, the multilevel approach will outperform the approach treating all effect sizes as independent, with a benefit that depends on the *average* number of effect sizes reported per study. Moreover, in our simulation design, we varied the size of the between-study variance of the treatment effect, but we assumed that this heterogeneity variance is equal for each outcome variable. We also varied the between-study covariance in the treatment effects—a positive covariance meaning that if, in a study, the treatment effect of an outcome is relatively large, the treatment effect of the other outcome variable also is likely to be large—but assumed a common covariation for all pairs of outcomes. These limitations can also explain why the three-level model performed even slightly better than the multivariate model when the effect was the same for all outcomes, although the multivariate model was used to generate the data: Whereas the multivariate model in which each outcome is regarded as a separate dependent variable includes a separate between-study variance parameter for each outcome and a separate between-study covariance parameter for each pair of outcomes, the three-level model assumes a common variance and covariance, resulting in a more parsimonious model to estimate with the same amount of data, resulting in more stable estimates. Whereas these assumptions are common in simulation studies (e.g., Hedges et al., 2010), they are often violated in practice. For instance, if two outcomes are measures of the same construct variable (e.g., two depression scales), the dependence is likely to be larger than for two outcomes referring to two different construct variables (e.g., a depression and an anxiety scale). Another situation where the covariation between outcomes can depend on the pair of outcomes is where studies include more than one (independent) sample and multiple outcomes were calculated for each sample: Outcomes from independent samples from the same study still are dependent, but less dependent than outcomes based on the same sample. We are currently investigating the performance of the proposed three-level approach by means of a simulation study set up in much the same way as the present study, for situations in which the number of outcomes varies over studies, if the between-study variance depends on the outcome, if the covariance between outcomes varies over pairs of outcomes, if samples sizes vary over studies, and if outcome effects are randomly sampled.

Author Note Wim Van den Noortgate, Faculty of Psychology and Educational Sciences, ITEC- IBBT, University of Leuven, Leuven, Belgium; José Antonio López-López, Faculty of Psychology, Universidad de Murcia, Spain; Fulgencio Marín-Martínez, Universidad de Murcia, Spain; and Julio Sánchez-Meca, Faculty of Psychology, Universidad de Murcia, Spain.

Appendix

Table 4 Bias (* 10,000) for the overall effect estimate for the multivariate, and the univariate three- and two-level models

			Common effect												Outcome-specific effect											
		$\sigma_{w_{1j}w_{1j'}}$	0				0.02				0.04				0				0.02				0.04			
		K	30		60		30		60		30		60		30		60		30		60		30		60	
J	$\sigma_{e_je_{j'}}$	N	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50
2	0	Multiv.	34	5	-4	1	-11	0	-11	10	8	-12	14	-5	18	12	-9	-10	-2	3	19	-7	-39	34	5	-4
		3-level	30	6	-7	0	-6	1	-10	12	7	-13	12	-5	15	3	-1	2	4	2	15	-3	-28	30	6	-7
		2-level	30	6	-7	0	-6	1	-10	12	7	-13	12	-5	15	3	-1	2	4	2	15	-3	-28	30	6	-7
	0.4	Multiv.	3	15	9	9	7	1	11	-2	-4	25	-8	1	19	0	-7	2	-9	-4	-6	3	-31	3	15	9
		3-level	2	16	10	11	11	-3	8	-2	-13	25	-6	3	23	-11	-3	4	-19	-3	9	5	-24	2	16	10
		2-level	2	16	10	11	11	-3	8	-2	-13	25	-6	3	23	-11	-3	4	-19	-3	9	5	-24	2	16	10
	0.8	Multiv.	11	-13	13	1	-5	-7	12	-4	-19	-3	-2	10	-44	31	-23	-12	-39	-27	-5	10	-54	11	-13	13
		3-level	10	-13	10	0	-2	-1	10	-5	-14	-7	-1	8	-15	3	-13	-1	-9	5	-2	10	-2	10	-13	10
		2-level	10	-13	10	0	-2	-1	10	-5	-14	-7	-1	8	-15	3	-13	-1	-9	5	-2	10	-2	10	-13	10
5	0	3-level	-36	-16	-41	-22	-52	-2	-45	-18	-25	-22	-33	-22	-50	-20	-45	-17	-43	-5	-41	-16	-53	-24	-48	-18
		2-level	-36	-16	-41	-22	-52	-2	-45	-18	-25	-22	-33	-22	-50	-20	-45	-17	-43	-5	-41	-16	-53	-24	-48	-18
	0.4	3-level	-52	-38	-47	-28	-27	-41	-34	-25	-31	-14	-38	-20	-47	-20	-40	-18	-58	-33	-42	-24	-43	-16	-38	-19
		2-level	-52	-38	-47	-28	-27	-41	-34	-25	-31	-14	-38	-20	-47	-20	-40	-18	-58	-33	-42	-24	-43	-16	-38	-19
	0.8	3-level	-36	-10	-39	-19	-47	-26	-29	-12	-27	-7	-25	-26	-38	-21	-42	-25	-55	-22	-43	-16	-37	-13	-37	-8
		2-level	-36	-10	-40	-19	-48	-26	-29	-12	-31	-7	-29	-26	-38	-21	-42	-25	-55	-22	-44	-16	-38	-13	-37	-8

Table 5 *MSE* (* 10,000) for the overall effect estimate for the multivariate and the univariate three- and two-level models

			Common effect												Outcome-specific effect											
		$\sigma_{w_{1j}w_{1j'}}$	0				0.02				0.04				0				0.02				0.04			
		K	30		60		30		60		30		60		30		60		30		60		30		60	
J	$\sigma_{e_je_{j'}}$	n	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50	25	50
2	0	Multiv.	33	26	15	13	37	28	18	14	40	35	19	16	49	43	24	20	59	48	27	23	65	60	32	29
		3-level	31	23	15	12	33	26	17	13	37	33	18	15	30	24	15	12	33	27	16	13	36	32	18	15
		2-level	31	23	15	12	33	26	17	13	37	33	18	15	30	24	15	12	33	27	16	13	36	32	18	15
	0.4	Multiv.	37	29	19	13	41	34	20	15	46	36	21	17	62	48	28	24	68	57	34	27	87	73	40	35
		3-level	35	27	18	13	39	32	20	14	44	34	21	16	36	26	18	14	38	30	20	15	42	36	22	16
		2-level	35	27	18	13	39	32	20	14	44	34	21	16	36	26	18	14	38	30	20	15	42	36	22	16
	0.8	Multiv.	45	32	23	15	47	36	23	17	51	40	25	19	75	55	35	26	88	71	41	30	110	92	50	41
		3-level	42	30	22	15	44	33	22	16	48	38	25	18	40	30	21	15	45	34	22	15	46	38	24	18
		2-level	42	30	22	15	44	33	22	16	48	38	25	18	40	30	21	15	45	34	22	15	46	38	24	18
5	0	3-level	11	10	6	4	17	15	9	7	21	21	11	10	12	9	6	5	16	15	8	7	23	20	12	10
		2-level	11	10	6	4	17	15	9	7	21	21	11	10	12	9	6	5	16	15	8	7	23	20	12	10
	0.4	3-level	20	14	10	7	25	20	12	10	29	25	15	12	20	14	10	7	25	19	13	10	31	25	15	12
		2-level	20	14	10	7	25	20	12	10	29	25	15	12	20	14	10	7	25	19	13	10	31	25	15	12
	0.8	3-level	29	19	14	9	34	25	17	12	40	28	19	14	29	19	15	9	33	25	17	12	40	29	19	13
		2-level	29	19	14	9	34	25	17	12	40	28	19	14	29	19	15	9	33	25	17	12	40	29	19	13

References

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

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

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Hoboken, NJ: Wiley.

Cheung, S. F., & Chan, D. K. S. (2008). Dependent correlations in meta-analysis: The case of heterogeneous dependence. *Educational and Psychological Measurement, 68,* 760–777.

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

Cooper, H. (2009). *Research synthesis and meta-analysis* (4th ed.). London: Sage.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7,* 177–188.

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.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 357–376). New York: The Russell Sage Foundation.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.

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

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

Higgins, J. P. T., Whitehead, A., Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine, 20,* 2219–2241.

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

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

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kreft, I. G. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York: Oxford University Press.

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.

López-lópez, J. A., Viechtbauer, W., Sánchez-Meca, J., & Marín-Martínez, F. (2010). *Comparing the performance of alternative statistical tests for moderators in mixed-effects meta-regression models*. Paper presented at the 5th Annual meeting of the Society for Research Synthesis Methodology. Cartagena, Spain.

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

Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H., & O'Mara, A. (2009). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research, 79,* 1290–1326.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13,* 85–116.

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. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 75–98.

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.

Shogren, K. A., Fagella-Luby, M. N., Bae, J. S., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior. *Journal of Positive Behavior Interventions, 6,* 228–237.

Snijders, T. A. B. (2003). Multilevel analysis. In M. Lewis-Beck, A. E. Bryman, & T. F. Liao (Eds.), *The SAGE Encyclopedia of Social Science Research Methods (Volume II)* (pp. 673–677). London: Sage.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage.

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

Thompson, S. G., Turner, R. M., & Warn, D. E. (2001). Multilevel models for meta-analysis, and their application to absolute risk differences. *Statistical Methods in Medical Research, 10,* 375–392.

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta-analysis. *Psychological Bulletin, 135,* 452–477.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental studies using hierarchical linear models. *School Psychology Quarterly, 18,* 325–346.

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

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2,* 142–151.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30,* 261–293.