Behavior Research Methods (2019) 51:1286–1304
https://doi.org/10.3758/s13428-018-1063-2

A demonstration and evaluation of the use of cross-classified random-effects models for meta-analysis

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

Published online: 5 June 2018

Abstract

It is common for the primary studies in meta-analyses to report multiple effect sizes, generating dependence among them. Hierarchical three-level models have been proposed as a means to deal with this dependency. Sometimes, however, dependency may be due to multiple random factors, and random factors are not necessarily nested, but rather may be crossed. For instance, effect sizes may belong to different studies, and, at the same time, effect sizes might represent the effects on different outcomes. Cross-classified random-effects models (CCREMs) can be used to model this nonhierarchical dependent structure. In this article, we explore by means of a simulation study the performance of CCREMs in comparison with the use of other meta-analytic models and estimation procedures, including the use of three- and two-level models and robust variance estimation. We also evaluated the performance of CCREMs when the underlying data were generated using a multivariate model. The results indicated that, whereas the quality of fixed-effect estimates is unaffected by any misspecification in the model, the standard error estimates of the mean effect size and of the moderator variables' effects, as well as the variance component estimates, are biased under some conditions. Applying CCREMs led to unbiased fixed-effect and variance component estimates, outperforming the other models. Even when a CCREM was not used to generate the data, applying the CCREM yielded sound parameter estimates and inferences.

Keywords Meta-analysis · Multiple effect sizes · Cross-classified random-effects model

In meta-analysis, researchers combine effect sizes (ESs) from multiple studies in order to obtain more accurate estimates of the parameters of interest, to increase statistical power in testing those parameters, and to study possible moderating effects of study characteristics. A common assumption in meta-analysis is the independence between ESs: It is assumed that (after accounting for moderating effects) the information given by one ES is completely unrelated to the information given by another ES. However, as many authors have already pointed out (e.g., Becker, 2000; Gleser & Olkin, 1994; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013, 2015), this assumption is often violated in practice, for multiple reasons. One of these reasons is that primary studies usually report more than one ES, often extracted from the same sample and/or from the same outcome variable. It is important that dependence between ESs within studies is detected and taken into account, because ignoring it can result in incorrect statistical inferences due to an underestimation of standard errors, which leads to an inflated Type I error rate, and thus an increased likelihood of false positives (Becker, 2000).

In general, three main approaches exist to deal with dependent ESs within studies: ignoring dependence, avoiding dependence, and modeling dependence. The first approach consists in considering ESs extracted from the same study as independent (as if each ES stemmed from an independent study). The second approach consists mainly in performing different meta-analyses for each type of ES reported in primary studies, selecting just one ES per study, or calculating the mean effect per study prior to combining the results in a meta-analysis. More information about the first two approaches can

✉ Belén Fernández-Castilla
belen.fernandezcastilla@kuleuven.be

[1] Faculty of Psychology and Educational Sciences, KU Leuven, University of Leuven, IICK Building, Box 1.33, Etienne Sabelaan 51, 8500 Kortrijk, Belgium

[2] Imec-ITEC, KU Leuven, University of Leuven, Leuven, Belgium

[3] Research Foundation Flanders (FWO), Brussels, Belgium

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

be found in Borenstein, Hedges, Higgins, and Rothstein (2009) and Becker (2000). Regarding the third approach (modeling the dependence between ESs), several techniques have been proposed over time. For instance, Becker (1992) proposed a multivariate general least squares approach to model dependence between the ESs from the same study. Other authors have also developed multivariate meta-analytic models in which the dependence between ESs is modeled assuming a fixed-effects, multivariate model (Raudenbush, Becker, & Kalaian, 1988) or assuming a random-effects multivariate two-level model (Kalaian & Raudenbush, 1996) in which variation over studies is also taken into account. The main disadvantage of these multivariate meta-analytic models is that researchers need to estimate in advance not only the sampling variance of the ESs, but also the sampling covariance, whereas the information required to estimate these covariances (e.g., the correlation between multiple outcome variables) is rarely reported in primary studies. Recently, Hedges, Tipton, and Johnson (2010) have proposed the robust variance estimation method (RVE), in which the dependence between ESs is handled by estimating robust standard errors of the parameters estimated. Implementation of this method requires some estimate of the correlation between the ESs, but the results are only affected to a minimal extent by the selected correlation value.

An alternative approach to handle dependence among ESs is through the use of multilevel modeling (Cheung, 2014; Raudenbush & Bryk, 2002; Van den Noortgate et al., 2013, 2015). When a standard random-effects meta-analytic model is fitted, two sources of variability are taken into consideration: the variability within studies (due to taking a sample of participants; Level 1) and the variability between studies (Level 2). An additional, intermediate level can be added to account for the presence of multiple ESs within studies, resulting in a three-level model. At the first level, the sampling variance is modeled: Each observed ES is considered equal to its true value plus a normally distributed random deviation. At the second level, the variation over the true ESs within studies is modeled (i.e., within-study variance). Finally, at a third level the between-studies variance is modeled, by defining study mean effects as randomly varying around an overall grand mean. Whereas the multilevel-modeling approach and the RVE method were shown to perform similarly (Moeyaert et al., 2016), applying multilevel models offers three important advantages relative to the application of the multivariate approach (Kalaian & Raudenbush, 1996; Raudenbush et al., 1988). First, multilevel models allow for modeling the correlation between ESs within studies without estimating the sampling covariances in advance. Second, by using random effects to model the differences between outcomes, we can make inferences to a larger population of outcomes, whereas with the multivariate approach the results apply only to the outcomes analyzed. The third advantage is derived from the previous one: Because in the multilevel approach outcomes are considered a random factor, the use of this approach is still appropriate if the outcomes that are studied differ greatly across studies. In contrast, the use of the multivariate approach in this situation may become unfeasible, because the number of ESs to be estimated and the number of between-outcome correlations that should be known before carrying out the analysis increase with the total number of outcomes.

Despite these advantages, some assumptions are made when multilevel models are applied to meta-analytic data: It is assumed that the correlation between each pair of outcomes is the same and that the between-study variance is the same for all outcomes. Therefore, the multivariate approach is preferable if the effects on more or less the same outcomes are reported across studies, if the researcher is interested in a separate estimate of the effect for each outcome (or in their comparison), and if the between-outcome covariances are available (which is rarely the case). In contrast, the application of a three-level model is especially advantageous in meta-analytic scenarios in which there are many different outcome variables (i.e., many different operationalizations), the presence of these variables varies widely across studies, or researchers want to generalize across outcome variables rather than focus on the specific outcome variables that were observed (e.g., Geeraert, Van den Noortgate, Grietens, & Onghena, 2004; Maes, Qualter, Vanhalst, Van den Noortgate, & Goossens, 2017). An additional advantage of using a three-level model is that by considering variability in ESs within studies as random effects, only one additional parameter (the between-outcome variance) has to be estimated, making it possible to include the characteristics of the outcomes as moderator variables to explain differences between the ESs within studies. Even when moderators are included, the (residual) within-study variance is modeled, which means that we do not assume that the ESs are the same within each category of the moderator. In contrast, introducing the characteristics of outcomes as predictors in a multivariate model is not possible, because the model already includes separate parameters for each outcome.

The three-level model can be extended by including additional levels. For instance, if studies are nested in countries or research teams, we can define an additional, upper level to handle more levels of dependence for study ESs. In the past, only hierarchical multilevel models have been proposed and used in the context of meta-analysis. Yet it may be the case that the meta-analytic data structure is not purely hierarchical. For instance, it is possible that ESs are nested within studies but also within the subscales that are used to measure an outcome (Fig. 1). These two factors are not nested within each other (multiple subscales can be used in a study, but at the same time multiple studies may use the same subscale), and hence ESs are cross-classified by study and subscale, which are then termed the "crossed factors" (CFs). To demonstrate this cross-classified structure, we will use a subset of data that

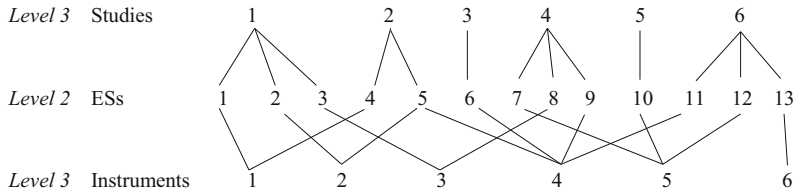


Fig. 1 Example of a meta-analytic cross-classified structure, using the graphical representation style of Fielding and Goldstein (2006). For simplicity, Level 1 (with one observed effect size [ES] per population ES) is not represented. Level 2 refers to the systematic variability between ESs that belong to the same study (and this variability is not due to the between-subscale nor to the between-study variance). At Level 3, the overall mean ES varies randomly over both studies and subscales

originate from the MASLO project, where the acronym stands for the *Meta-Analytic Study of Loneliness* (Maes, Van den Noortgate, Fustolo-Gunnink, et al., 2017), in which gender differences in loneliness are explored.[1] There are 57 studies that include 68 standardized mean differences. In the dataset, included in Appendix Table 7, two variables are reported in addition to the study identification number, each ES value (Hedges's *g*) and their sampling variance: the "Outcome" variable represents the types of outcomes reported across studies, all of them representing the loneliness construct, and the "Subscale" variable refers to the scale or subscale used to measure loneliness. In this dataset, seven possible subscales were used, including the UCLA Loneliness Scale (code 1), the Social dimension of the SELSA (code 2), the Emotional dimension of the RTLS (code 5), and so on.

To tackle this meta-analysis, a researcher might consider the application of different models, such as a multivariate or a hierarchical three-level model. Let us say that, in this specific case, the researcher is interested in gender differences in loneliness, rather than in separate estimates of the gender difference for each outcome. Therefore, a multivariate approach is less useful in this case. Moreover, the use of the multivariate approach would require prior estimates of the covariance between ESs within each study, but primary studies do not always report information on the correlations between outcomes that is needed to estimate these sampling covariances. A three-level meta-analytic model could be applied, in which the sampling variance is set at the first level, the within-study variance is modeled at the second level and the between-studies variance is modeled at third level. Nevertheless, each subscale for measuring loneliness might have its own effect, since each subscale has its own characteristics (e.g., length, style in the formulation of the questions, specific facet of loneliness on which they focus, etc.), which can influence the value of the ESs. To take a potential influence of the "Susbcale" variable into consideration, one option is to model the subscale effects by means of a moderator variable with fixed effects. In this way, we would estimate the effect for each of these seven specific subscales. However, it might be the case that there are not enough observations within each subscale category to accurately estimate the subscale effects (i.e., in this example, there is only one observation of Subscale 3; see Appx. A), so treating subscales as a random variable could be a solution for this potential estimation problem (Snijders & Bosker, 2012). If, moreover, the interest of the researcher is not in the specific subscales, but rather the researcher wants to draw inferences about the general gender differences, and possibly investigate how these gender differences vary over subscales, considering "Subscale" as a random variable might be more appropriate. The theoretical implication of treating "Subscale" as a random variable is that the seven subscales are seen as a sample of a population of subscales, making it possible to generalize the results of the meta-analysis to all existing subscales. Following this reasoning, the initial three-level hierarchical structure turns into a cross-classified structure in which subscales and studies constitute two crossed random factors at the third level, as outcomes belong to a specific study and at the same time outcomes are measured using a certain subscale. In other words, at Level 3, ESs have some variation due to (1) the variability between studies and (2) the variability between subscales, whereas at Level 2 some systematic within-study variability in ESs (not due to the between-study or between-subscale variability) might exist. This type of structures can be analyzed with an adapted version of the so called *cross-classified random-effects model* (CCREM; Fielding & Goldstein, 2006; Goldstein, 1994; Rasbash & Goldstein, 1994), and studies and subscales would constitute two CFs at Level 3.

Table 1 contains the results when the data were analyzed with a three-level CCREM versus a hierarchical three-level model (recognizing the clustering of ESs within studies and ignoring the variation between subscales). The estimate of the standard error of the pooled ESs is almost three times smaller when a hierarchical three-level model rather than the cross-classified model is fitted.[2] The direct consequence of this difference is that the combined ES is statistically significant when a hierarchical multilevel model is applied, whereas the combined ES is not significant when a CCREM is applied.

[1] For this example, we will analyze just a subset of the studies published until 2013, in which the sample concerned only elderly people.

[2] Note that in this subset there are seven subscales, and seven units might not be enough to accurately estimate the between-subscale variance. In the complete dataset there are 27 subscales. When analyses were performed on the complete dataset, the standard error estimate obtained with a hierarchical three-level model was still three times smaller (.008) than the estimate obtained after applyting a CCREM (.024).

Table 1 Empirical demonstration of the effects of ignoring a crossed factor at the third level, with meta-analytic data from the MASLO project

	CCREM		HLM	
	Estimate (*SE*)	CI	Estimate (*SE*)	CI
Intercept (combined ES)	– .038 (.084)	(– .245, .169)	– .052* (.026)	(– .104, – .0001)
Between-ESs variance	.009		.032	
Between-study variance	.012		.000	
Between-subscale variance	.036		–	

* $p < .05$; ES = effect size; *SE* = standard error; CI = 95 % confidence interval; HLM = hierarchical linear model

Therefore, the conclusion would drastically change as a function of the model that is fitted. In this case, it can be seen that there is some variation between subscales (even more than variation between studies), which explains why ignoring that variability has an impact on the inferences regarding the mean effect.

In meta-analytic data, a cross-classified structure can occur for several reasons and at different levels. For instance, in De Wit, Greer, and Jehn's (2012) meta-analysis, the authors investigated the relationship between intragroup conflict and group outcomes. In several primary studies, multiple ESs were reported for sets of group outcomes (including, e.g., cohesion, trust, satisfaction, commitment, and identification). Hence, as in the previous illustration, three levels could be distinguished in the data with a cross-classification of study and outcome at the third level. The model for this cross-classified meta-analytic data structure is the following:

$$d_{j(mk)} + \delta_{000} + r_{j(mk)} + w_{(mk)} + v_m + u_k \quad (1)$$

where each type of residual, $r_{j(mk)}$, $w_{(mk)}$, v_m, and u_k, is assumed to be independently and normally distributed with a mean of zero and variance σ_r^2, σ_w^2, σ_v^2, and σ_u^2, respectively, and the sampling variance σ_r^2 is assumed to be known. At Level 1, the sampling error variability is modeled as $d_{j(mk)}$, where the j^{th} ES estimate for outcome m and study k depends on the sample. At Level 2, the model captures the variability within combinations of studies and outcomes (σ_w^2). Finally, at Level 3, variability in the ESs is modeled as a function of the outcome (cross-classified factor residual, v_m) and study (cross-classified factor residual, u_k), where δ_{000} represents the grand mean or the combined ES. If the random component v_m or $u_k$were deleted, the model would become a hierarchical three-level model. As we stated before, if the types of outcomes reported across studies (e.g., cohesion, trust, satisfaction, commitment, and identification) were limited, if the researcher was interested in estimating a separate ES for each type of outcome, and if the covariance between ESs were known, a multivariate approach could also be applied.

Another type of cross-classified meta-analytic data can be found in the meta-analysis conducted by Gilboa, Shirom, Fried, and Cooper (2008), in which the correlation between job stressors (measured through many different outcomes, including role ambiguity, role conflict, role overload, work-family conflict, job insecurity, and environmental uncertainty) and job performance (measured in different ways; e.g., self-reported, by supervisors, objective ratings, and qualitative performance) was synthesized. In this case, each ES within a study could represent the correlation between job performance and each of a wide range of job stressors (e.g., role conflict), and at the same time, each correlation (independently of the outcome or the job stressor used) could be influenced by the way in which job performance was rated (e.g., rated by supervisors). The types of job stressors and the way of evaluating job performance are not nested within each other: The same job stressor might have been correlated with different ratings of job performance, and the same way of rating job performance might have been used for calculating the correlation with multiple job stressors. The different types of job stressors (outcomes) and the different ways of evaluating job performance (ratings), therefore constitute two cross-classified factors at the second level, which are at the same time nested within studies (Level 3), meaning that both the type of job stressor and the way of rating job performance differ from study to study. As in the previous example, if (1) the number of different outcomes was limited, (2) the researchers were interested in the effect for each outcome, (3) more or less the same outcomes were reported across studies, and (4) the covariance among ESs could be estimated, a multivariate approach could have been applied. Also, if there were limited categories of outcomes or of ratings, any of those variables could be introduced as a fixed moderator variable in a hierarchical three-level analysis. However, because the number of categories is relatively high, introducing outcomes or ratings as fixed moderators might lead to estimation problems, since some of those categories might refer to just one ES. Furthermore, the researcher might not be interested in the specific outcome or rating effect, but might just want to account for the possible variability that these variables could generate, and/or want to explain the variation by looking at the relation between the outcome or rating characteristics and the size of the effect. In these cases, applying a CCREM with

outcomes and ratings as CFs at the second level would be more appropriate. The statistical model of this example is as follows:

$$d_{(ml)k} + \delta_{000} + r_{(ml)k} + w_{mk} + v_{lk} + u_k \quad (2)$$

where, as in the previous example, residuals are typically assumed to be independently and normally distributed with zero means and variances of σ_r^2, σ_w^2, σ_v^2, and σ_u^2, and the sampling variance, σ_r^2, is assumed to be known. At Level 1, an ES for outcome m (CF1), evaluated with rating l (CF2) from study k, varies randomly due to a known sampling variance (σ_r^2). At Level 2, the cross-classified effects of outcome (w_{mk}) and rating (v_{lk}) are modeled, and, finally, at Level 3, the between-study variability is modeled (σ_u^2). As in hierarchical multilevel models, moderator variables can be included in any of the levels of the model. For instance, Eq. 2 can be extended to include characteristics of the first CF (e.g., Z_{mk}), the second CF (e.g., X_{lk}), or the studies (e.g., W_k) as predictors. Writing the whole model in one equation, the inclusion of these moderator variables results in Eq. 3:

$$d_{(jm)k} + \delta_{000} + \gamma_1 Z_{mk} + \gamma_2 X_{lk} + \gamma_3 W_k + w_{mk} + v_{lk} + u_k + r_{(ml)k} \quad (3)$$

As is shown, there is a wide variety of potential cross-classified structures in the context of meta-analysis, especially on those occasions for which there are many different outcomes, subscales, ratings, subpopulations, and so forth, within or across studies, and the researcher is not interested in the ES for each separate category of those variables. Recognizing and correctly specifying these models may be highly relevant, since several simulation studies (including, e.g., Luo & Kwok, 2009, and Meyers & Beretvas, 2006) have already shown the undesirable consequences of misspecifying a cross-classified structure when conventional raw data are analyzed. Specifically, Meyers and Beretvas found that when the cross-classified structure was ignored, the standard errors associated with the incorrectly modeled variables were underestimated and both the variance of the nonignored CF and the variance at the first level of the model were overestimated. Luo and Kwok extended this work to a three-level model and explored the effects of the misspecification when the cross-classified structure was ignored at the second or the third level. The results were similar to those found by Meyers and Beretvas for the standard errors, but with respect to the variance components, the variance of the nonignored CF was underestimated while the variances at adjacent levels were overestimated. These results correspond with the finding that when the variance at a certain level in a conventional hierarchical model is ignored, the variance is then apportioned into a variance at adjacent levels (Berkhof & Kampen, 2004; Moerbeek, 2004; Opdenakker & Van Damme, 2000; Tranmer & Steel, 2001; Van den Noortgate, Opdenakker, & Onghena, 2005). Thus, the incorrect specification of a cross-classified raw data structure leads to incorrect estimates of the variance components at different levels, as well as an inflated Type I error rate due to the underestimation of the standard errors for fixed-effect parameter estimates.

The aims of this article are to introduce the use of CCREMs in the context of meta-analysis, to explore their performance, and to study the statistical consequences of model misspecification. The results of the real-data analysis (Table 1) suggest the importance of accounting for all random effects, but the exact influence of misspecification cannot be derived directly from previous research on the use of CCREMs outside the meta-analytic context. In meta-analysis, the variance at the first level (i.e., sampling variance) is fixed for each ES estimate and is assumed to be known, which means that the consequences of misspecifying a cross-classified meta-analytic data structure might differ from those found and described by Meyers and Beretvas (2006) and by Luo and Kwok (2009), because the variance associated with an ignored CF cannot be displaced or added to the variance at the first level of the model (for more information about variance component estimation when raw data are analyzed, see Searle, Casella, & McCulloch, 1992). In addition to comparing results when using the meta-analytic CCREM versus a conventional three-level or two-level meta-analytic model, we will also compare the estimates obtained when using the RVE method (Hedges et al., 2010) for handling within-study dependence in ESs. A second simulation study will be set up to study the robustness of the approach by describing the consequences of applying a CCREM when the underlying meta-analytic data are actually generated using a three-level multivariate model, with data generated for multiple outcomes (as multiple dependent variables), and with random effects within studies (e.g., subscales, rating, . . .) and across studies.

Finally, although different cross-classified data structures can be encountered in the context of meta-analysis (as we illustrated in the previous paragraphs), we will now focus on three-level data with a cross-classification at the second level (matching the second example described earlier).

Simulation study 1

The first simulation study was conducted to investigate estimation of CCREMs for meta-analysis and to explore how the estimates of the fixed effects and the variance components were affected when one or both CFs were ignored or when a statistical technique that did not model the cross-classification directly was applied.

Method

Data generation and analysis ESs were simulated to fit the CCREM defined in Eq. 3. These ESs can represent

standardized mean differences, Fisher's z-transformed correlations, log odds ratios, or any ES measure with an approximately normal distribution from which it is possible to estimate the sampling variance. In this case, the ES chosen for simulation was the standardized mean difference, as it is one of the most common ESs used by applied researchers. The simulated datasets included three dichotomous or dummy moderator variables: two variables at the second level (one related to each of the CFs) and one dummy variable at the third level, as meta-analysts commonly explore the effects of moderator variables on ESs. In addition, we took into account that in reality, not all levels of both CFs may be available within each study. For instance, in the meta-analysis of Gilboa et al. (2008), it could be the case that within one study, the authors used only one way of evaluating job performance (e.g., supervisor-rated performance), but there could be correlations of job performance with different job stressors (e.g., work–family conflict and job insecurity). In the same way (but inverting the CFs), it could be the case that within one study, just one job stressor was reported (e.g., work–family conflict) but that stressor could have been correlated with different operationalizations of job performance (e.g., supervisor-rated performance, self-report performance and qualitative performance). In other words, it is likely that the ESs within one study belong to just one level of CF1, whereas they belong to several levels of CF2, or the other way around. Therefore, we varied the number of levels of the CFs over studies within the meta-analysis in order to simulate more realistic data: In one third of the studies, the ESs belonged to a unique level of CF1, whereas they were nested in several levels of CF2 (e.g., only one outcome variable has been measured, and job performance has been rated through multiple ratings). In the second third of the studies, the ESs belonged to several levels of CF1 but were nested in just one level of CF2 (e.g., several outcome variables have been measured, and just one rating has been used to evaluate job performance), and in the last third of the studies, the ESs were nested in different levels of both CFs (e.g., with ESs representing several outcomes and ESs representing several ways of rating job performance).

Effect sizes were then generated according to four scenarios, differing from each other in whether the number of reported ESs was the same for all primary studies (balanced–unbalanced condition) and whether the moderator variables at the second level were correlated (uncorrelated–correlated condition). We simulated these scenarios because simulating simplified balanced data can help in understanding the results, whereas simulating unbalanced data might reflect more realistic situations. Furthermore, we wanted to be able to generalize the results to more complex situations, such as the case in which moderator variables correlate.

In the condition in which moderator variables were correlated at the second level, ESs were disproportionally distributed over the values of the dummy variables, meaning that about 70% of the ESs had the values 1 and 1 or 0 and 0 for the two dummy variables and about 30% of the ESs had values of 0 and 1, or vice versa, for the dummy variables at the second level. For generating this situation, twice the ESs were generated as in the uncorrelated-covariate condition, and then 70% of the ESs belonging to value 0 in one dummy variable and to value 1 in the other dummy variable, and vice versa, were removed, whereas only 30% of the ESs belonging to values 0 and 0 or 1 and 1 of both dummy variables were deleted. In this way, the expected Cramer's V index between the values of the moderator variables to be approximately .40. In the uncorrelated condition the two covariates were independent, so the ESs were approximately equally distributed over the four combinations of values of the two dummy variables.

In the balanced and uncorrelated condition, 24 ESs per study were generated. In the balanced and correlated condition, 48 ESs per study were generated, so that after deleting 50% of the ESs, as described above, the expected number of ESs per study was 24 as well. Each of the ESs referred to one out of four levels of CF1 (m = 1, 2, 3, 4), and to one out of three levels of a second CF (CF2 l = 1, 2, 3), and these levels were allowed to differ from study to study. Then, many of these ESs were deleted to generate the cross-classified structure explained in the previous paragraph: in one third of the studies ESs belonged to just one level of CF1 (so the ESs belonging to Level 2, 3, and 4 of CF1 were deleted), in another third, ESs belonged to just one level of CF2 (so the ESs belonging to Levels 2 and 3 of CF2 were removed), and, finally, in the last third of studies, the ESs belonged to several levels of both CFs (70% of the ESs were randomly removed). In the end, each study contained approximately the same number of ESs (more or less ±7). Both CFs were allowed to vary randomly across studies, meaning that we assumed that each ES—referring to one level of CF1 and other level of CF2—had been randomly sampled from a population of possible levels of CF1 and CF2. Finally, CF1 and CF2 were purely nested within studies.

In the unbalanced condition, 48 or 96 ESs per study were generated, depending on whether the covariates at the second level were correlated (then 96 ESs per study were generated) or uncorrelated (then 48 ESs were generated). Each of the ESs belonged to one out of six levels of CF1 (m = 1, 2, 3, 4, 5, 6) and to one out of four levels of CF2 (l = 1, 2, 3, 4), and these levels were allowed to differ from study to study. Once the ESs were obtained, and after deleting many of them in order to create the cross-classification mentioned in the previous sections, 10% of the ESs were randomly deleted for one third of the studies, for another third of the studies 50% of the ESs were randomly deleted and for the final third of studies, 90% of the ESs were randomly deleted. This means that overall, about 50% of the ESs were deleted (in addition to all those ESs that were already deleted to create the desired cross-classification and to generate correlation between moderator

variables). In the end, the expected total number of ESs in this condition were the same as in the balanced condition. The only difference is that in this case, there were different numbers of ESs across studies. For example, Study 1 could include ten ESs, while Study 11 could include just one ES.

These two conditions were crossed, resulting in four scenarios. In the first, *balanced–uncorrelated* scenario, there were approximately the same numbers of ESs within studies, and the moderator variables related to the characteristic of the ESs within studies were not correlated. In the *balanced–correlated* condition, there was the same number of ESs in all studies, but the moderator variables that reflected the characteristics of each ES within studies were correlated (i.e., the presence of a value of 1 in Moderator Variable 1 was related to the appearance of a value of 1 in Moderator Variable 2). In the *unbalanced–uncorrelated* condition, there were different numbers of ESs within studies and the moderator variables were not correlated. Finally, in the *unbalanced–correlated* condition, there were also different numbers of ESs in each study, and the moderator variables expressing characteristics of each ES within studies were correlated.

Afterward, the ESs in all conditions were analyzed according to five models (Fig. 2). Model 1 (Eq. 3) is the CCREM model used to simulate the data. Models 2 and 3 ignored one of the CFs—CF2 (v_{lk}) or CF1 (w_{mk}), respectively—resulting in a hierarchical three-level model. In Model 4, a standard three-level model was fitted in which both CFs were ignored but the variation between ESs within the same study was still modeled at Level 2 (without differentiating between CF1 and CF2). In Model 5, a two-level model was fitted, resulting in a standard random-effects model in which only the variation across studies was modeled. Within this last model, parameters were estimated in two different ways: using restricted maximum likelihood (REML; later on referred to as *Model 5a*) or using the RVE method proposed by Hedges et al. (2010) to provide robust standard errors for the fixed-effect parameter estimates (*Model 5b*).

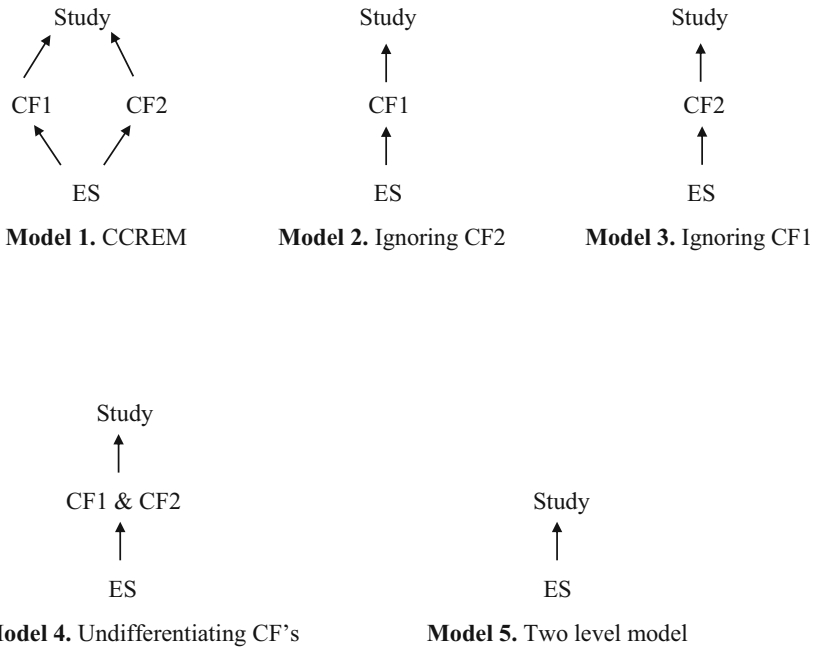
Three moderator variables were included in each of these models. The moderator variable W_k referred to a dummy-coded study characteristic, and the moderator variables Z_{mk} and X_{lk} referred to dummy-coded characteristics of CF1 and CF2, respectively. The regression coefficients for each of these moderator variables are represented using γ_3, γ_1, and γ_2 respectively.

Conditions Various conditions were simulated within the four previously described scenarios (balanced or unbalanced ESs, correlated or uncorrelated covariates). According to the review of Rubio-Aparicio, Marín-Martínez, Sánchez-Meca, and López-López (2017), the median number of studies included in a meta-analysis is 25. Therefore, we simulated two quantities around this number: the number of studies (k) was equal to either 18 or 36. The sampling variances used to generate the data were equal to .10 and .05, which correspond to the sampling variances of the standardized mean differences with two groups of size 40 or 80. The variance at the first level was fixed at this sampling variance. For determining realistic values of the variances of the residuals across the levels of the CF1, CF2, and studies, we looked at several published meta-analyses that had used three-level models (e.g., Geeraert et al., 2004; Van den Bussche, Van den Noortgate, & Reynvoet, 2009). In these studies, the variances at the second and third levels ranged from .02 to .12. To mimic realistic data, values of .05, .10, and .15 were selected for each of the three sources of variance. The combined ES (δ_{000}, grand mean) was set to 0 (null effect) or to .5 (moderate effect,) and the moderator effects were set to 0 (null effect) or .3 (moderate effect). All these combined conditions resulted in 432 different conditions in each of the four scenarios. For each combination (432 conditions), we simulated and analyzed 1,000 datasets, so in total 4 × 432 × 1,000 = 1,728,000 datasets were simulated and analyzed.

SAS was used to generate and analyze the data. All models, except for Model 5b, were estimated with the PROC MIXED procedure implemented in SAS (the SAS code is given in Appx. B), which uses the REML method for estimating the variances. For applying the RVE method (Model 5b), the R code reported by Hedges et al. (2010) was translated into SAS language.[3]

Evaluation of the simulation results The fixed- and random-effect parameter estimates from the five models were summarized across the 1,000 iterations for each condition. To evaluate the estimations of the fixed-effect parameters, bias was approximated by subtracting the true value from the mean estimated value. The mean squared error (*MSE*) was also calculated in order to evaluate the accuracy of the estimates in each model. For the variance components, the bias was evaluated by subtracting the true value from the median estimated value and, for the standard errors, bias was evaluated by comparing the median of the estimated standard errors in a certain condition with the standard deviation of the fixed-effect estimates in that condition. It can be assumed that the standard deviation of the fixed-effect estimates is a fairly good approximation of the true standard deviation of the sampling distribution of the estimator, because we simulated a large number of datasets for each combination of parameters. We preferred to use the median rather than the mean of the standard errors and variance components because the distributions of these parameters tend to be skewed. Moreover, we looked at the coverage percentages of the 95% confidence intervals (CIs) in each condition. The standard error of a 95% CI is approximately .007, so an appropriate coverage percentage would be expected to be between 94% and 96%. Appropriate coverage

[3] This SAS code is available upon request from the first author.



Model 1. CCREM **Model 2.** Ignoring CF2 **Model 3.** Ignoring CF1

Model 4. Undifferentiating CF's **Model 5.** Two level model

Fig. 2 Illustration of the models that will be fitted in Simulation Study 1

of the intervals would indicate a lack of bias in the parameter estimates and the corresponding standard errors.

Afterward, the relative bias (RB) of each parameter estimated (fixed effects, standard errors, and variance components) was approximated by dividing the estimated bias by its true value (Hoogland & Boomsma, 1998). For evaluating the values of the RB, we followed the cutoffs proposed by Hoogland and Boomsma: An RB was moderate but acceptable if its value was between 5% and 10%, whereas an RB was considered unacceptable if its value was above 10%.

Finally, analyses of variance (ANOVAs) were conducted in order to detect which conditions were related to a higher or a lower bias. To focus our discussion on the most influential design factors, only effects with a partial eta-squared (η_p^2) higher than .02 are discussed.

Results

An important first result is that we found no bias in the estimates of the combined ES (i.e., the estimates of the intercept), nor in the estimates of the regression weights for the moderator variables at any level, condition, or model, and therefore these estimates are not further discussed here. In contrast, the *MSE*s of the fixed effects and the estimates of the standard errors and variance components varied over conditions and models. The factors that mainly predicted the bias in the standard error and variance estimates were the true values of the variance components (.05, .10, .15), the four condition categories (correlated–uncorrelated covariates and equal–unequal numbers of ESs within studies), and the number of studies within the meta-analysis (18 or 36), but we found no effect of the value of the overall intercept, the moderator effects, or the number of participants within studies.

We will present the results first for estimates of the fixed-effect parameters (*MSE*, standard errors, and coverage of the CI) and then for estimates of the variance components. Afterward, the results of the second simulation will be discussed in a similar way, though more briefly.

Fixed effects

Intercept The standard error of the combined ES was properly estimated in all conditions when a CCREM was applied (RB below 5%), but for the other models the estimates were biased under some circumstances. The conditions that mainly moderated bias were the type of model fitted ($\eta_p^2 = .74$) the number of ESs in the studies within the meta-analysis ($\eta_p^2 = .09$), the correlation between the covariates at the second level ($\eta_p^2 = .19$), and the number of studies within the meta-analysis ($\eta_p^2 = .30$). When one of the CFs was ignored, the standard error of the combined ES was on average underestimated (with an average RB of – 12.9% for Model 2 and – 9.9% for Model 3). In Table 2 we can see that this underestimation became worse (1) as the total variance of the model increased (so the variance of the ignored CF also increased), (2) as the number of ESs within studies varied, and (3) as the moderator variables were correlated at the second level. Notice that in Model 3 (in which CF1 is ignored) the RB of the standard error of the combined ES was within the acceptable thresholds when the number of ESs within studies was constant and when there were different numbers of ESs within studies but the moderator variables were

Table 2 Percentages of relative bias (RB; in percentages) of the standard errors and coverage pecentages of the 95% confidence intervals (CIs) for the combined effect size (intercept)

Variance	Number of ESs	Relation of Moderators	*k*	RBs by Model						95% CI Coverage Percentages by Model					
				1	2	3	4	5a	5b	1	2	3	4	5a	5b
.05	Balanced	Uncorrelated	18	– 4	– 12	– 9	– 11	– 19	– 3	94	92	93	92	90	95
			36	– 1	– 9	– 6	– 8	– 16	– 1	95	93	93	93	90	95
		Correlated	18	– 4	– 12	– 10	– 13	– 22	– 11	95	92	93	92	88	93
			36	– 1	– 9	– 7	– 11	– 20	– 6	95	93	93	92	89	94
	Unbalanced	Uncorrelated	18	– 2	– 11	– 9	– 12	– 20	– 6	95	93	93	92	89	94
			36	– 2	– 12	– 9	– 13	– 20	– 5	94	92	92	91	88	94
		Correlated	18	– 3	– 12	– 11	– 15	– 23	– 15	94	92	93	91	88	92
			36	– 2	– 11	– 10	– 15	– 23	– 10	95	92	93	91	87	93
.15	Balanced	Uncorrelated	18	– 2	– 13	– 8	– 11	– 21	– 3	95	92	93	92	89	95
			36	– 1	– 12	– 7	– 10	– 20	– 3	95	92	93	92	89	94
		Correlated	18	– 3	– 15	– 12	– 16	– 28	– 12	95	91	93	91	86	92
			36	0	– 12	– 10	– 15	– 27	– 7	95	92	93	91	86	93
	Unbalanced	Uncorrelated	18	– 3	– 14	– 11	– 16	– 25	– 7	95	91	93	91	87	94
			36	– 1	– 13	– 10	– 15	– 24	– 5	95	91	93	91	87	94
		Correlated	18	– 3	– 15	– 13	– 19	– 29	– 16	95	91	91	89	85	91
			36	– 2	– 14	– 12	– 19	– 29	– 10	95	91	92	89	84	93

RB = relative bias (in percent); *k* = number of studies within the meta-analysis. Model 1 = CCREM; Model 2 = three-level model in which Crossed Factor 2 (CF2) is ignored; Model 3 = three-level model in which Crossed Factor 1 (CF1) is ignored; Model 4 = three-level model in which CFs are undifferentiated; Model 5a = two-level model; Model 5b = two-level model using RVE method for correcting standard errors

uncorrelated. The standard error of the synthesized ES was also underestimated when the CFs were not distinguished with Model 4 (on average, RB = −14.7%), and especially when a standard two-level random-effects model (Model 5a) was fitted (RB = −24.3% RB). In both Models 4 and 5a, the RB increased as the total variance of the model increased, if the covariates were correlated, and if there were different number of ESs across studies. When the RVE method was applied (Model 5b), the RB was on average within the recommendable thresholds (RB = −7.3%). However, if we look at Table 2 we can see that the RB varied as a function of the number of studies included in the meta-analysis and the correlation between covariates at the second level: When there were only 18 studies within the meta-analysis and when the moderator variables were correlated, the standard error of the intercept was underestimated.

A second way to evaluate the fixed-parameter and standard error estimates is by looking at the coverage proportions of the CI estimates. The coverage proportions (Table 2) for the 95% CIs of Models 2, 3, 4, and especially 5a were too small (the coverage proportion was in general below .94). Given the unbiased point estimates, this confirms that the standard errors were too small. In line with previous results, when the RVE method was applied, the coverage proportions for the 95% CIs were approximately good, except when moderator variables were correlated and there were only 18 studies in the meta-analysis. When a CCREM was applied, all the coverage proportions were between the expected appropriate values (.94–.96).

Finally, with respect to the precision of the estimates, the *MSE*s of the estimates obtained through the CCREM were lower than the *MSE*s for other models (Table 3). The estimates obtained through the RVE method had the highest *MSE*, especially when the number of ESs varied over studies.

Moderator variable at the study level (γ_3) On average, the standard errors of the study characteristic moderator variables were estimated without substantial bias (i.e., RB was below 5%), even if one of the CFs was ignored in the model (Models 2 and 3), the CFs were modeled together at an intermediate level (Model 4), or a standard random-effects model was fitted with standard errors, either adjusted or not adjusted (Models 5a and 5b). There was no effect of the balanced–unbalanced distinction or of the correlated–uncorrelated distinction.

Moderator variables at the intermediate CF factors' level (γ_1 and γ_2) The main predictor of the bias of the standard errors for the γ_1 and γ_2 moderator coefficient estimates was the type of model fitted ($\eta_p^2 = .84$). The estimates of the standard errors of the covariates at the second level were unbiased when the correct CCREM was applied (i.e., RB below 5%; Table 4). When one of the CFs was ignored, the standard error of the coefficient for the moderator associated with the ignored CF

Table 3 Mean squared error for the fixed effects by condition and model

	Fixed Effect	Balanced ESs		Unbalanced ESs	
		Uncorrelated	Correlated	Uncorrelated	Correlated
Model 1	δ_{000}	.0316	.0402	.0364	.0406
	γ_3	.0367	.0384	.0396	.0396
	γ_1	.0130	.0198	.0155	.0191
	γ_2	.0135	.0201	.0162	.0195
Model 2	δ_{000}	.0337	.0430	.0396	.0440
	γ_3	.0379	.0396	.0408	.0407
	γ_1	.0138	.0215	.0165	.0208
	γ_2	.0144	.0227	.0190	.0233
Model 3	δ_{000}	.0328	.0423	.0386	.0432
	γ_3	.0376	.0395	.0407	.0407
	γ_1	.0137	.0221	.0177	.0221
	γ_2	.0142	.0217	.0171	.0212
Model 4	δ_{000}	.0339	.0451	.0403	.0469
	γ_3	.0374	.0395	.0405	.0407
	γ_1	.0140	.0226	.0177	.0227
	γ_2	.0148	.0232	.0190	.0237
Model 5a	δ_{000}	.0341	.0457	.0409	.0477
	γ_3	.0374	.0395	.0408	.0409
	γ_1	.0141	.0229	.0179	.0229
	γ_2	.0151	.0237	.0194	.0242
Model 5b	δ_{000}	.0364	.0542	.0515	.0604
	γ_3	.0376	.0398	.0416	.0419
	γ_1	.0190	.0347	.0302	.0402
	γ_2	.0201	.0336	.0299	.0383

δ_{000} = combined effect size; γ_3 = moderator variable at the study level; γ_1 = moderator variable describing Crossed Factor 1 (CF1); γ_2 = moderator variable describing Crossed Factor 2 (CF2); Model 1 = CCREM; Model 2 = three-level model in which CF2 is ignored; Model 3 = three-level model in which CF1 is ignored; Model 4 = three-level model in which CFs are undifferentiated; Model 5a = two-level model; Model 5b = two-level model using RVE method for correcting standard errors

was underestimated. For example, when CF2 was not recognized in the model (Model 2), the standard error of the variable related to that factor [$SE(\hat{\gamma}_2)$] was underestimated, whereas the standard error of the variable related to CF1 [$SE(\hat{\gamma}_1)$] remained within the acceptable RB thresholds. The same happened when CF1 was eliminated from the model (Model 3): The standard error of the coefficient for the moderator describing CF1 [$SE(\hat{\gamma}_1)$] was underestimated, whereas the RB of the included factor [$SE(\hat{\gamma}_2)$] was below 10%.

Table 4 shows the RBs of the standard errors of the coefficients for the CF1 moderator variable (γ_1) and the coverage proportions of the 95% CI. These patterns are mirrored for the standard errors of the CF2 moderator variable (γ_2) (not shown in the table). The ANOVA showed that the three factors that determined the amount of bias in the CI coverage rates estimated using Models 2 and 3 were: the correlation between the covariates ($\eta_p^2 = .18$), the balanced/unbalanced number of ESs within studies in the meta-analysis ($\eta_p^2 = .35$), the number of studies ($\eta_p^2 = .67$), and above all the amount of variance that was ignored ($\eta_p^2 = .86$). When the ignored CF had a large variance (.15), the underestimation of the stan dard error of the covariate associated with the missing CF reached, on average, an RB of − 58.9%. In addition, if the numbers of ESs within the studies that constituted the meta-analysis were unequal and the moderator variables were uncorrelated, the RB was even higher (RB = −60.8%). When both CFs were modeled as a single source of variance (Model 4), and especially when a two-level model was fitted (Model 5a), the standard errors of both the moderator variables were also underestimated. This underestimation was worse when there were different numbers of ESs per study. The RBs of the estimates obtained through the RVE method were within the accept able cutoffs (i.e., RB below 10%), except when all of the following were true: The moderators were correlated, there were different numbers of ESs per study, and the total

number of studies included in the meta-analysis was low. In these cases, the RB equaled or exceeded the acceptable value.

The results for coverage proportions for the 95% CI were very similar when the residuals for CF1 were eliminated (Model 3) and when a two-level model was fitted (Model 5a), in that the coverage proportions were extremely low (see Table 4). In the case in which CFs were undifferentiated (Model 4), the coverage proportions were also too low. When the correct CCREM was applied (Model 1), when the moderator did not describe the ignored CF2 (Model 2), and when the RVE method was used to handle within-study dependence, the actual coverage proportions were close to the nominal value (between .94 and .96).

Finally, regarding the precision of the estimates (Table 3), the lowest *MSE* value was obtained when the correct CCREM was applied, whereas the largest value (and thus the least precise method) was obtained with the RVE method.

Variance components

Variance between studies The between-study variance was highly overestimated in all models except for the correct CCREM. A one-way ANOVA showed that in general the overestimation of the between-study variance only depended on the amount of variance that was ignored in each model and on the true value of the between-study variance (Table 5). When one of the CFs was removed from the model (Models 2 and 3), the variance between studies was on average overestimated by 64.1% and 55.1%, respectively, and this overestimation became worse in scenarios in which the variance of the ignored CF was larger. When both CFs were undistinguished (Model 4) and when a two-level model was fitted (Model 5a), the study variance was also overestimated (on average, RBs of 113.3% and 134.3% for Models 4 and 5a, respectively). The highest overestimation of the between-study variance occurred when the RVE method was used (Model 5b; on average, this RB was 243.7%), although on average the estimated between-study variance equaled the sum of all the variabilities in the data (e.g., if the variance between the CFs was .10 and the between-study variance was .05, the estimated between-study variance with the RVE method was on average .25, which corresponds with the sum of all of the variances). For all models (except for Model 1), the higher the true value of the between-study variance, the less biased the estimates.

Variance of the CFs The variances of the CFs were accurately estimated under all conditions when a CCREM was applied. When one of the CFs was missing in the model (Models 2 and 3), the CF that stayed in the model was slightly overestimated. This overestimation depended only on the amount of variance attributed to both CFs (the one that was ignored and the one that was kept in the model). As can be seen in Table 6, when

Table 4 Percentages of relative bias (RB; in percentages) of the standard error and coverage proportions of the 95% confidence intervals (CIs) of the moderator variable describing Crossed Factor 1 (CF1; γ_1)

Variance	Number of ESs	Relation of Moderators	*k*	RB by Model						95% CI Coverage Proportions by Model					
				1	2	3	4	5a	5b	1	2	3	4	5a	5b
.05	Balanced	Uncorrelated	18	– 2	– 2	– 35	– 19	– 36	2	.95	.95	.80	.89	.77	.96
			36	– 2	– 2	– 36	– 19	– 38	0	.95	.95	.78	.88	.77	.95
		Correlated	18	– 5	– 7	– 34	– 21	– 38	– 6	.95	.95	.81	.89	.78	.95
			36	– 2	– 4	– 32	– 19	– 36	– 4	.95	.95	.81	.89	.78	.94
	Unbalanced	Uncorrelated	18	– 2	– 2	– 40	– 26	– 41	– 6	.95	.95	.75	.85	.74	.95
			36	– 3	– 3	– 41	– 27	– 43	– 4	.95	.95	.75	.85	.73	.95
		Correlated	18	– 4	– 5	– 38	– 27	– 42	– 10	.95	.94	.77	.86	.75	.94
			36	– 2	– 3	– 38	– 26	– 42	– 8	.95	.95	.77	.86	.74	.94
.15	Balanced	Uncorrelated	18	– 2	– 2	– 56	– 27	– 58	2	.95	.95	.61	.85	.59	.96
			36	0	0	– 55	– 25	– 58	0	.95	.95	.62	.86	.59	.95
		Correlated	18	– 4	– 7	– 54	– 29	– 58	– 8	.95	.94	.64	.85	.60	.94
			36	– 2	– 5	– 53	– 27	– 57	– 6	.95	.94	.64	.85	.60	.94
	Unbalanced	Uncorrelated	18	– 2	– 2	– 60	– 36	– 62	– 5	.95	.95	.56	.80	.55	.96
			36	0	0	– 61	– 35	– 62	– 4	.95	.95	.56	.80	.54	.95
		Correlated	18	– 3	– 5	– 58	– 35	– 61	– 11	.95	.94	.59	.81	.56	.94
			36	– 3	– 5	– 60	– 36	– 62	– 8	.95	.94	.57	.79	.53	.94

k = number of studies within the meta-analysis; Model 1 = CCREM; Model 2 = three-level model in which CF2 is ignored; Model 3 = three-level model in which CF1 is ignored; Model 4 = three-level model in which CFs are undifferentiated; Model 5a = two-level model; Model 5b = two-level model using RVE method for correcting standard errors

Table 5 Percentage relative bias of the between-study variance

Study True Variance	CFs' True Variance	Estimating Model					
		Model 1	Model 2	Model 3	Model 4	Model 5a	Model 5b
.05							
	.05	– 6	54	46	95	112	204
	.10	– 8	114	98	194	228	407
	.15	– 11	173	151	292	345	611
.10							
	.05	– 5	25	21	46	54	101
	.10	– 5	55	46	96	113	202
	.15	– 7	84	73	145	171	303
.15							
	.05	– 4	15	13	30	35	66
	.10	– 5	34	29	62	74	133
	.15	– 5	55	46	96	113	202

CFs = crossed factors; Model 1 = CCREM; Model 2 = three-level model in which CF2 is ignored; Model 3 = three-level model in which CF1 is ignored; Model 4 = three-level model in which CFs are undifferentiated; Model 5a = two-level model; Model 5b = two-level model using RVE method for correcting standard errors

CF2 was ignored (Model 2), the variance of CF1 was overestimated if the ignored variance was high, but also if the actual variance of CF1 was really low. A corresponding pattern was found for Model 3.

Simulation Study 2

The second simulation study aimed to check the consequences of applying a CCREM when the underlying meta-analytic data were generated to fit a three-level multivariate model.

Method

Data generation and analysis For this simulation, a multivariate three-level model was used to generate the data (Appx. C). At the first level, participants' scores (i.e., the raw data) were generated for three different outcomes, and the membership of each individual in either the control (0) or the treatment (1) group was indicated by a dummy variable (X). At the second level, the intercepts and the treatment effects for each of the three outcomes were allowed to vary randomly over a random factor with four levels (e.g., subscales), and finally, at the third level, the intercepts and the treatment effects were allowed to vary randomly across studies. We generated raw data instead of ESs directly because when a multivariate model is applied, the covariance between the ESs needs to be known, and for estimating these covariances the raw data are needed (or studies should report estimates of the covariances, which often is not the case). After the generation of raw data, standardized mean differences and their sampling (co)variances were estimated following the formulas proposed by Gleser and Olkin (1994) for multiple-endpoint studies. A multivariate three-level model was applied to the ESs, following the approach described by Kalaian and Raudenbush (1996). A CCREM, with two crossed random factors at the intermediate level (e.g., outcomes and subscales), was also applied to the ESs. The number of ESs was the same in all studies (i.e., balanced), and no moderator variables were generated.

Table 6 Percentages of relative bias of the cross-classified factor variance estimates

	Model and Variance of CF1					
	Model 2			Model 3		
Variance of CF2	.05	.10	.15	.05	.10	.15
.05	10	4	3	9	23	37
.10	23	11	7	3	11	18
.15	36	18	12	1	6	11

CF1 = Crossed Factor 1; CF2 = Crossed Factor 2; Model 2 = three-level model in which CF2 is ignored, Model 3 = three-level model in which CF1 is ignored

Conditions The conditions were the same in terms of the numbers of studies, group sizes, and mean ESs (in this case, the value of the mean treatment effect was the same as the mean ES). The first-level variances were fixed at 1, so that the covariance between outcomes equaled the correlation, which could be 0 or .5 (for conditions in which the outcomes, respectively, did not covary vs. covaried substantially). At the second and third levels, the intercepts and the treatment variances were all equal to .05 or .10. The covariance between each pair of outcomes' intercept variances was generated to be .05 at

both the second and third levels, and the covariances between each pair of outcomes' slopes was generated to be .02, also at both levels. The intercept–slope covariances were generated using a value of − .025. These values for the covariances were chosen because the correlation coefficients at upper levels are likely to be small (Lipsey & Wilson, 2001). The mean intercepts were fixed to 0. All of this resulted in 2 × 2 × 2 × 2 = 16 conditions, and 1,000 datasets were generated in each of the 16 conditions, resulting in 16,000 datasets.

Evaluation of the simulation results The bias and RB were calculated separately for each outcome when the multivariate model was estimated. However, since we estimated a CCREM model that provided one overall mean ES estimate, the estimates of the ESs for each of the three outcomes from the multivariate model were also averaged and then compared to the mean estimate calculated using the CCREM. Regarding the variance components, the three between-study and between-subscale variances obtained with the multivariate approach (one for each outcome) were averaged and then summed in order to know the average total variance in the model.

Results

As in the previous simulation, the estimates of the combined ES were unbiased for both the CCREM and the multivariate model. In the multivariate approach, both the estimates of the effects for the three individual outcomes and the estimates of the overall effect were unbiased (average RB = – 3.6%). Similar results were obtained for the CCREM, in which the RB of the pooled ES was even lower (on average, RB = – 1.1%). Regarding the standard errors, in the multivariate approach the standard errors of the three separated outcomes were correctly estimated, with an RB below – 4.8%. When a CCREM was applied, the standard error estimate of the combined ES was also accurate (overall RB = – 2.0%). As for the variance components, the total expected variance in the model was .10 and .20 for the conditions in which both variance components (i.e., studies and subscales) equaled .05 and .10, respectively. In the condition in which the covariance among ESs was substantial (i.e., .50) and the total variance was .10, the total estimated variance of the CCREM was closer to the real total variance than was the total estimated variance for the multivariate model (.099 as compared to .086), whereas when the true total variance was expected to be .20, both models underestimated the total amount of variance, although this time the estimates were more similar across models: .168 for the CCREM and .175 for the multivariate approach. Finally, when there was no covariation among ESs, both models led to similar estimates when the true total variance was .10 (.080 for CCREM and .087 for the multivariate model), but when the total amount of variance was larger (.20) the CCREM led to too-low estimates of the total amount of variance (.150), although the multivariate approach also underestimated the total true value (.178).

Discussion

In this study, we have proposed for the first time the use of a cross-classified random-effects model for analyzing meta-analytic data with a cross-classified structure, and we have illustrated, through one empirical and three theoretical examples, how cross-classification structures can be present in meta-analytic data. In the first simulation study, we have shown the consequences that several kinds of misspecification have on the variance and fixed-effect estimates when ESs are cross-classified at the second level. The simulation study varied several factors in order to mimic a variety of situations that can occur, including scenarios in which the ESs within studies belong to several levels of one CF but to just one level of the other CF, in which the number of ESs varies widely across studies, and in which the moderator variables are correlated at the second level. Even though some studies have already explored the consequences of ignoring cross-classification structures (Luo & Kwok, 2009; Meyers & Beretvas, 2006) or explored the repercussions of ignoring full levels in a hierarchical structure (Berkhof & Kampen, 2004; Moerbeek, 2004; Opdenakker & Van Damme, 2000; Tranmer & Steel, 2001; Van den Noortgate et al., 2005), as far as we know, no previous simulation studies have explored the performance of the CCREM in the context of meta-analysis in which the variance at the first level is known in advance. One first conclusion of this study is that applying CCREM when the data structure is cross-classified results in unbiased estimates of the fixed effects and variance components, even if there are different numbers of ESs across studies, and even if the moderator variables at the intermediate level are correlated.

We have also explored what would happen if the researcher just considered one of the CFs but ignored the other (Models 2 and 3). As was mentioned in Luo and Kwok (2009), ignoring one CF at the intermediate level has consequences similar to ignoring a level of nesting: The ignored variance is divided over the adjacent levels. In our simulation, when one of the CFs was removed from the model (Models 2 and 3), the between-study variance and the variance of the CF that remained in the model were overestimated (this relative overestimation became substantial if the variance of the CF that stayed in the model was low). These results do not agree with the findings of Luo and Kwok, in which if a CF was ignored at the second level, the variance moved toward the adjacent levels, and the variance of the CF that stayed in the model was underestimated. One plausible reason why here the CF that stayed in the model was overestimated is that the variance of the ignored CF could not go downward, because the Level

1 residuals' variance was fixed. Thus, the ignored variance mainly went upward (toward the study level) but part of the variance stayed at the same level, where it is added to the other CF variance (in line also with the results of Meyers & Beretvas, 2006). Regarding the standard errors of the fixed effects, the standard error of the moderator describing the missing CF was underestimated, whereas the standard error of the variable associated with the CF that remained in the model was well-estimated. These results are in line with the results obtained both by Meyers and Beretvas and by Luo and Kwok, in which a variable related to an ignored CF was modeled as a Level 1 (instead of a Level 2) variable, and therefore its standard error was underestimated. Neither Luo and Kwok nor Meyers and Beretvas included information about the estimation of the standard error of the intercept, but in the study by Van den Noortgate et al. (2005), the authors showed through a simulation study that the standard error of the intercept was underestimated when the variance at the top level was removed (in their case, the top level was the fourth level), but not when the second or the third level was ignored. Contrary to these results, in our simulation study we found that the standard error of the intercept was underestimated if one CF was ignored, even when the CFs were indistinguishable or when the intermediate level was omitted, especially when (1) the number of studies included in the meta-analysis was small, (2) the ignored variance was high, (3) covariates were correlated, and (4) there were different numbers of ESs within studies.

When a three-level model was fitted and the CFs were not explicitly distinguished at the second level, the variance at the study (third) level was again overestimated. The standard error of the synthesized ES was underestimated, and this underestimation got worse if the total variance of the three different sources (study, CF1, and CF2) increased, if the covariates were correlated, and if there were different numbers of ESs across studies. The standard errors of both moderator variables at the second level were underestimated, and this underestimation was worse when there were different numbers of ESs per study. The consequences of fitting a standard random-effects model—overestimation of the study variance and underestimation of the standard errors of all fixed effects in the model (intercept and coefficients for all moderator variables)—were the same in direction as the consequences of fitting a three-level model without differentiating between the CFs included, but were worse in magnitude.

Finally, using the RVE method led to an inflated estimate of the between-study variance, although this estimate accurately reflected the total amount of variance in the data. Contrary to the results obtained by Moeyaert et al. (2017), our results showed that the performance of the RVE method was affected by the number of studies within the meta-analysis and by the correlation of moderator variables at the intermediate level. When there were only 18 studies and the covariates were correlated, the standard error of the combined ES was underestimated. Regarding the standard error of the second-level covariates, the estimates were within the acceptable range, except when the covariates were correlated, when there were different numbers of ESs in the studies, and when the total number of studies included in the meta-analysis was low. In addition, the RVE method resulted in the largest *MSE*s for all fixed effects, demonstrating less accuracy in the estimates. In the study by Moeyaert et al., no important differences between methods (multilevel modeling and RVE) were found in the estimation of the standard error of the combined ES. A plausible reason for our obtaining different results is that Moeyaert et al. simulated conditions in which 25 or 50 studies were included in the meta-analysis and there were similar numbers of ESs across studies (a balanced design). In our simulation, we considered the situations in which the studies included were fewer in number (k = 18) and in which the ESs were unbalanced across studies (plus possible correlation between the moderators), and these differences could easily account for the discrepancy in the results. On the basis of the results of our simulation study, we recommend using the CCREM rather than the RVE method when a small number of studies are included in the meta-analysis and when the ESs are unbalanced.

It should be mentioned that the standard error estimate of the moderator effect of the study characteristic was always accurate across the five models. These results are in line with the results obtained by Moerbeek (2004): Ignoring intermediate levels (e.g., the second level) does not have an effect on the estimation of the standard error of a moderator variable at the top level (e.g., the third level). In our study, five models ignored the variance at the second level either partially (i.e., Models 2, 3, and 4) or fully (i.e., Models 5a and 5b), but that did not affect the estimation of the standard error of the moderator variable at the top level.

Regarding the second simulation study, we have shown that even if a CCREM is applied to a three-level multivariate model, it is still possible to obtain accurate estimates of the fixed effects and their standard errors, with resulting correct inferences about the results of the meta-analysis. Only in the condition in which the covariance among ESs was zero and the total variance in the model was high did the multivariate approach result in better estimates of the total variance, although this variance was still underestimated. In the rest of the conditions, the performance of the CCREM was similar to that of the multivariate approach, even though the meta-analytic data were generated to fit a multivariate model and *not* the CCREM. In addition, not even the application of the correct model (i.e., a three-level multivariate model) led to accurate estimates of the variance components, even though the structure of the data was relatively simple in terms of balance and the model lacked covariates. These findings support the conclusion that the CCREM performs well under a variety of conditions. Moreover, the application of CCREMs

has the advantage that the covariances among ESs do not need to be known in advance and that a meta-analysis is still feasible if different sets of outcomes, subscales, ratings, or any other variable are measured across studies.

In summary, the results of this study show how the inappropriate modeling of cross-classified ESs within studies distorts both estimation of the variance components and estimation of the standard errors of the fixed effects. Specifically, the improper estimation of the variance components is related to the amount of variance ignored, whereas the inadequate estimation of the standard errors depends on the balanced/unbalanced number of ESs within studies, on the correlation between the moderator variables, on the amount of variance ignored, and on the number of studies within the meta-analysis. Although this simulation study has focused on a specific structure (the three-level model with cross-classification at the intermediate level), we also would expect negative consequences from a misspecification of the cross-classified structure at other levels. Therefore, when planning a multilevel meta-analysis, we recommend careful consideration of the underlying data structure in order to guarantee appropriate estimates of the combined ESs, standard errors, and variance components, with special consideration for any potential cross-classifications.

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

Appendix A

Table 7 Data and code for the Maes, Van den Noortgate, Fustolo-Gunnink, et al. (2017) example

Study	Outcome	Subscale	*g*	Variance	Precision
1	1	1	– .251	.024	41.455
2	1	1	– .069	.001	1,361.067
3	1	5	.138	.001	957.620
4	1	1	– .754	.085	11.809
5	1	1	– .228	.020	49.598
6	1	6	– .212	.004	246.180
6	2	7	.219	.004	246.095
7	1	1	.000	.012	83.367
8	1	2	– .103	.006	162.778
8	2	3	.138	.006	162.612
8	3	4	– .387	.006	160.133
9	1	1	– .032	.023	44.415
10	1	5	– .020	.058	17.110
11	1	1	.128	.017	59.999
12	1	1	– .262	.032	31.505
13	1	1	– .046	.071	14.080
14	1	6	– .324	.003	381.620
14	2	6	– .409	.003	378.611
14	3	7	.080	.003	386.319
14	4	7	– .140	.003	385.542
15	1	1	.311	.005	185.364
16	1	1	.036	.005	205.063
17	1	6	– .259	.001	925.643
17	2	7	.196	.001	928.897
18	1	1	.157	.013	74.094
19	1	1	.000	.056	17.985
20	1	1	.000	.074	13.600
21	1	6	– .013	.039	25.425
21	2	7	– .004	.039	25.426
22	1	1	– .202	.001	1,487.992
23	1	1	.000	.086	11.628
24	1	1	– .221	.001	713.110
25	1	1	– .099	.001	749.964
26	1	5	– .165	.000	6,505.024
27	1	1	– .523	.063	15.856
28	1	1	.000	.001	1,611.801
29	1	6	.377	.045	22.045
29	2	7	.575	.046	21.677
30	1	1	.590	.074	13.477
31	1	1	.020	.001	1,335.991
32	1	1	.121	.043	23.489
33	1	1	– .101	.003	363.163
34	1	1	– .101	.003	369.507
35	1	1	– .104	.004	255.507
36	1	1	– .270	.003	340.761
37	1	1	.179	.150	6.645
38	1	2	.468	.020	51.255
38	2	4	– .479	.020	51.193
39	1	5	– .081	.024	42.536
40	1	1	– .071	.043	23.519
41	1	1	.201	.077	13.036
42	1	6	– .070	.006	180.844
42	2	7	.190	.006	180.168
43	1	1	.277	.013	79.220
44	1	5	– .086	.001	903.924
45	1	5	– .338	.002	469.260
46	1	1	.262	.003	290.330
47	1	5	.000	.003	304.959
48	1	1	– .645	.055	18.192
49	1	5	– .120	.002	461.802
50	1	5	– .286	.009	106.189
51	1	1	– .124	.006	172.261

Table 7 (continued)

Study	Outcome	Subscale	*g*	Variance	Precision
52	1	1	.023	.028	35.941
53	1	5	− .064	.001	944.600
54	1	1	.000	.043	23.010
55	1	1	.000	.014	72.723
56	1	5	.000	.012	85.832
57	1	1	.000	.012	85.832

The "Outcome" variable enumerates the number of outcomes within each study. The "Subscale" variable represents the subscale used to measure loneliness for each outcome, for example, code 1 refers to the UCLA scale, and code 5 refers to RTLS scale. The SAS code for a CCREM with studies (CF1) and subscales (CF2) cross-classified at the third level is as follows:

```
proc mixed data=Loneliness;
    class study subscale outcome;
    weight precision;
    model ES= /SOLUTION ddfm=sat;
    Random intercept /SUB=study;
    Random intercept /SUB=subscale;
    Random intercept /SUB=outcome(study
    *subscale);
    parms .5 .5 .5 1/hold = (4);
    run;
```

Three random effects are defined using the `random` statement: a random effect that varies over studies, one that varies over subscales, and one that varies over outcomes (which are nested within a cross-classification of studies and subscales). The `parms` statement is used to give starting values for the variance estimates. The `hold` option is used to indicate that all these starting values, except the fourth one can be updated. By fixing the last variance (which is the residual variance, the variance at Level 1) and using the inverse of the sampling variance as a weight, we ensure that the variance for each observed effect size equals its sampling variance.

The code follows for a hierarchical three-level model that does not differentiate between CF1 and CF2:

```
proc mixed data=Loneliness;
    class study outcome;
    weight precision;
    model ES= /SOLUTION ddfm=sat;
    Random intercept /SUB=study;
    Random intercept /SUB= outcome(study);
    arms .5 .5 1/hold = (3);
    run;
```

Appendix B

SAS codes of the models fitted in the first simulation study

Model 1: CCREM (with two random factors nested within studies)

```
proc mixed data= data;
     class study CF1 CF2;
     weight precision;
     model d= dummy1 dummy2 dummy3/SOLUTION ddfm=sat;
     Random intercept /SUB=study;
     Random intercept /SUB=CF2(study);
     Random intercept /SUB=CF1(study);
     parms .1 .5 .5 1/hold = (4);
     run;
```

Model 2: Three-level model ignoring CF2

```
proc mixed data= data;
     class study CF1;
     weight precision;
     model d= dummy1 dummy2 dummy3/SOLUTION ddfm=sat;
     Random intercept /SUB=study;
     Random intercept /SUB=CF1(study);
     parms .5 .5 1/hold = (3);
      run;
```

Model 3: Three-level model ignoring CF1

```
proc mixed data= data;
     class study CF2;
     weight precision;
     model d= dummy1 dummy2 dummy3/SOLUTION ddfm=sat;
     Random intercept /SUB=study;
     Random intercept /SUB=CF2(study);
     parms .5 .5 1/hold = (3);
     run;
```

Model 4: Three-level model undifferentiating CF1 and CF2

```
proc mixed data= data;
     class study ESs;
     weight precision;
     model d= dummy1 dummy2 dummy3/SOLUTION ddfm=sat ;
     Random intercept /SUB=study;
     Random intercept /SUB= ESs(study);
     parms .5 .5 1/hold = (3);
     run;
```

Model 5a: Standard random effects model (2-level)

```
proc mixed data= data;
     class study;
     weight precision;
     model d= dummy1 dummy2 dummy3/SOLUTION ddfm=sat ;
     Random intercept /SUB=study;
     parms .5 1/hold = (2);
     run;
```

The multivariate three-level model used to generate the data in the second simulation study

At the first level, the model has three different equations, one for each of three outcome variables. Y_{i1lk} refers to the score of subject i on Outcome 1 on subscale l (l = 1, 2, 3, 4) within study k. The subjects can belong either to the control (X = 0) or the treatment (X = 1) group, which is represented by a dummy variable X. The treatment effect is represented by β_1 and the expected control group score by β_0. The indices for both coefficients indicate that both the control group performance and the size of the effect can depend on the subscale and the study,

$$\begin{cases} Y_{i1lk} = \beta_{01lk} + \beta_{11lk}X_{ilk} + e_{i1lk} \\ Y_{i2lk} = \beta_{02lk} + \beta_{12lk}X_{ilk} + e_{i2lk} \\ Y_{i3lk} = \beta_{03lk} + \beta_{13lk}X_{ilk} + e_{i3lk} \end{cases},$$

where the sampling errors are distributed following a multivariate normal distribution:

$$\begin{bmatrix} e_{i1lk} \\ e_{i2lk} \\ e_{i3lk} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{e_1} & & \\ \sigma_{e_1e_2} & \sigma^2_{e_2} & \\ \sigma_{e_1e_3} & \sigma_{e_2e_3} & \sigma^2_{e_3} \end{bmatrix} \right)$$

At the second level, the intercepts (β_0) and the treatment effects (β_1) are allowed to vary randomly across subscales, around a mean value for study k (the θs).

$$\begin{cases} \beta_{01lk} = \theta_{01k} + v_{01lk} \\ \beta_{02lk} = \theta_{02k} + v_{02lk} \\ \beta_{03lk} = \theta_{03k} + v_{03lk} \end{cases} \qquad \begin{cases} \beta_{11lk} = \theta_{11k} + v_{11lk} \\ \beta_{12lk} = \theta_{12k} + v_{12lk} \\ \beta_{13lk} = \theta_{13k} + v_{13lk} \end{cases}$$

where the intercept's random residuals (v_{01lk}, v_{02lk} , v_{03lk}) and the treatment effects residuals (v_{11lk}, v_{12lk}, v_{13lk}) for the three outcomes within subscales l and study k are distributed following a multivariate normal distribution:

$$\begin{bmatrix} v_{01lk} \\ v_{02lk} \\ v_{03lk} \\ v_{11lk} \\ v_{12lk} \\ v_{13lk} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{v_{01}} & & & \\ \sigma_{v_{01}v_{02}} & \sigma^2_{v_{02}} & & \\ \vdots & \vdots & \ddots & \\ \sigma_{v_{01}v_{13}} & \sigma_{v_{02}v_{13}} & \cdots & \sigma^2_{v_{13}} \end{bmatrix} \right)$$

Finally, at the third level, the intercepts (θ_0) and treatment effects (θ_1) are allowed to vary across studies, around mean values γ_0 and γ_1.

$$\begin{cases} \theta_{01k} = \gamma_{01} + u_{01k} \\ \theta_{02k} = \gamma_{02} + u_{02k} \\ \theta_{03k} = \gamma_{03} + u_{03k} \end{cases} \qquad \begin{cases} \theta_{11k} = \gamma_{11} + u_{11k} \\ \theta_{12k} = \gamma_{12} + u_{12k} \\ \theta_{13k} = \gamma_{13} + u_{13k} \end{cases}$$

where the intercepts' random study effects (u_{01k}, u_{02k}, u_{03k}) and the treatment effects (u_{11k}, u_{12k}, u_{13k}) for the three outcomes are distributed following a multivariate normal distribution:

$$\begin{bmatrix} u_{01k} \\ u_{02k} \\ u_{03k} \\ u_{11k} \\ u_{12k} \\ u_{13k} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{u_{01}} & & & \\ \sigma_{u_{01}u_{02}} & \sigma^2_{u_{02}} & & \\ \vdots & \vdots & \ddots & \\ \sigma_{u_{01}u_{13}} & \sigma_{u_{02}u_{13}} & \cdots & \sigma^2_{u_{13}} \end{bmatrix} \right)$$

References

Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational and Behavioral Statistics*, *17*, 341–362. https://doi.org/10.3102/10769986017004341

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.

Berkhof, J., & Kampen, J. K. (2004). Asymptotic effect of misspecification in the random part of the multilevel model. *Journal of Educational and Behavioral Statistics*, *29*, 201–218. https://doi.org/10.3102/10769986029002201

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

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*, 211–229. https://doi.org/10.1037/a0032968

De Wit, F. R., Greer, L. L., & Jehn, K. A. (2012). The paradox of intragroup conflict: A meta-analysis. *Journal of Applied Psychology*, *97*, 360–390.

Fielding, A., & Goldstein, H. (2006). Cross-classified and multiple membership structures in multilevel models: An introduction and review (Research Report No. 791). University of Birmingham, Department of Eduction and Skills. ISBN 1 84478797 2.

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.

Gilboa, S., Shirom, A., Fried, Y., & Cooper, C. (2008). A meta-analysis of work demand stressors and job performance: Examining main and moderating effects. *Personnel Psychology*, *61*, 227–271. https://doi.org/10.1111/j.1744-6570.2008.00113.x

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 339–355). New York, NY: Russell Sage Foundation.

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods and Research*, *22*, 364–375.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. https://doi.org/10.1002/jrsm.5

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

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*, 227–235. https://doi.org/10.1037/1082-989X.1.2.227

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

Luo, W., & Kwok, O. M. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral*

Research, *44*, 182–212. https://doi.org/10.1080/00273170902794214

Maes, M., Qualter, P., Vanhalst, J., Van den Noortgate, W., & Goossens, L. (2017). *Gender differences in loneliness: A meta-analysis*. Leuven, Belgium: Unpublished manuscript, KU Leuven.

Maes, M., Van den Noortgate, W., Fustolo-Gunnink, S. F., Rassart, J., Luyckx, K., & Goossens, L. (2017). Loneliness in children and adolescents with chronic physical conditions: A meta-analysis. *Journal of Pediatric Psychology, 42*, 622–635.

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, *41*, 473–497. https://doi.org/10.1207/s15327906mbr4104_3

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, *39*, 129–149. https://doi.org/10.1207/s15327906mbr3901_5.

Moeyaert, M., Ugille, M., Beretvas, S. N., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, *20*, 559–572. https://doi.org/10.1080/13645579.2016.1252189

Opdenakker, M. C., & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: an illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, *11*, 103–130.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, *19*, 337–350.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*, 111–120. https://doi.org/10.1037/0033-2909.103.1.111

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

Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., & López-López, J. A. (2017). A methodological review of meta-analyses of the effectiveness of clinical psychology treatments. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-017-0973-8

Searle, S., Casella, G., & McCulloch, C. E. (1992). Variance components. New York, NY: Wiley.

Snijders, T., & Bosker, R. (2012). Multilevel analysis: An introduction to basic and applied multilevel analysis (2nd ed.). London, UK: Sage.

Tranmer, M., & Steel, D. G. (2001). Ignoring a level in a multilevel model: Evidence from UK census data. *Environment and Planning A*, *33*, 941–948.

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: a meta-analysis. *Psychological Bulletin*, *135*, 452–477. https://doi.org/10.1037/a0015329

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*, 576–594. https://doi.org/10.3758/s13428-012-0261-6

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, *47*, 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

Van den Noortgate, W., Opdenakker, M. C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, *16*, 281–303.