Behavior Research Methods (2020) 52:2031–2052
https://doi.org/10.3758/s13428-020-01373-9

The application of meta-analytic (multi-level) models with multiple random effects: A systematic review

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

Published online: 11 March 2020

Abstract

In meta-analysis, study participants are nested within studies, leading to a multilevel data structure. The traditional random effects model can be considered as a model with a random study effect, but additional random effects can be added in order to account for dependent effects sizes within or across studies. The goal of this systematic review is three-fold. First, we will describe how multilevel models with multiple random effects (i.e., hierarchical three-, four-, five-level models and cross-classified random effects models) are applied in meta-analysis. Second, we will illustrate how in some specific three-level meta-analyses, a more sophisticated model could have been used to deal with additional dependencies in the data. Third and last, we will describe the distribution of the characteristics of multilevel meta-analyses (e.g., distribution of the number of outcomes across studies or which dependencies are typically modeled) so that future simulation studies can simulate more realistic conditions. Results showed that four- or five-level or cross-classified random effects models are not often used although they might account better for the meta-analytic data structure of the analyzed datasets. Also, we found that the simulation studies done on multilevel meta-analysis with multiple random factors could have used more realistic simulation factor conditions. The implications of these results are discussed, and further suggestions are given.

Keywords Systematic review · meta-analysis · multiple effect sizes · multilevel models

In any scientific discipline, it is common to find units that are nested in higher-level clusters. An example in educational research is the nesting of children in classrooms. Children from the same classroom are exposed to common stimuli that might make their behavior in general more alike than the behavior of children from different classrooms. Examples of clustered data structures in biology or medicine are animals clustered in phylogenetic families, or patients nested within hospitals. This nesting of observations within higher-level clusters involves the possible existence of dependency among observations. That is, each observation does not give unique information, and not taking this interdependency into account may lead to the overestimation of the available information and hence to inflated Type I error rates. Multilevel models are methods that appropriately account for dependency among observations (Hox, 2002).

One application of multilevel models is meta-analysis (Hox, 2002; Raudenbush & Bryk, 2002). Meta-analysis refers to the set of statistical tools that enable the combination of evidence from different studies that address the same research question (Glass, 1976). Once effect sizes are extracted from primary studies, they are pooled together by applying a fixed or random effects model. In a fixed effect model, it is assumed that there is only one underlying population effect size, and that the observed effect sizes deviate from this population effect due to sampling variation only. A random effects model assumes that each study has its own population effect, that is, that effect sizes vary due to sampling variation and also due to systematic differences that exist across studies. Under this model, not only the combined effect size is estimated but also the variance of the overall effect across studies. Often, it is also of the interest of the meta-analyst to find (moderator) variables that explain the variation of effect sizes across studies.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-020-01373-9) contains supplementary material, which is available to authorized users.

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

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

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

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

Raudenbush and Bryk (1985) proposed the use of multilevel models for performing meta-analysis. Meta-analytic data indeed has a hierarchical structure: study participants are nested within studies. Because raw data are rarely available, effect sizes that summarize the raw data from the study participants are often used. Raudenbush and Bryk (1985) differentiated between the within-study model and the between-study model. In multilevel notation, the within-study model refers to the variation of effect sizes due to sampling variation at level 1:

$$d_k = \gamma_k + r_k \tag{1}$$

where d_k represents the effect size reported in study k, γ_k refers to the population effect size of study k, and r_k is a normally distributed random residual with mean 0 and variance $\sigma^2_{r_k}$. This is the sampling variance, that for commonly used effect sizes can be calculated before the meta-analysis, and therefore is assumed to be known. The between-studies model models the variation of effect sizes across studies, and constitutes the second level in the multilevel model:

$$\gamma_k = \delta_0 + u_k \tag{2}$$

where δ_0 is the combined effect size, and u_k a random residual with expected value zero. Typically, the study residuals are assumed to follow a normal distribution with variance σ^2_u, which therefore represents the between-studies variance. The model can also be written in one single equation:

$$d_k = \delta_0 + u_k + r_k \tag{3}$$

This two-level model is equivalent to the traditional random effects model (Hedges & Olkin, 1985), meaning that all random effects meta-analyses are, by nature, multilevel models (Pastor & Lazowski, 2018). The random effects model can be extended by including predictor variables in an attempt to explain the heterogeneity between studies, that is, to find variables that moderate the effect. Note that if the between-studies variance equals zero, then the random effects model reduces to a fixed effect model.

Multilevel models have been proven to be as effective and accurate to estimate the parameters of interest in meta-analysis as other traditional random effects approaches, such as DerSimonian and Laird's (1986) or Hedges and Olkin's (1985) method, with the additional advantage that multilevel models are more flexible (Van den Noortgate & Onghena, 2003). For instance, multiple predictors can be easily incorporated in the model to explain heterogeneity among effect sizes across studies. Furthermore, additional random effects can be added in the meta-analytic model for dealing with various kinds of dependency among effect sizes within and between studies.

In primary studies, it is quite common to report multiple effect sizes because, for instance, the construct of interest is measured using different instruments, and/or several treatment groups are compared to a common control group, and/or different sub-samples are used. The effect sizes that stem from the same study are likely to be more similar, because they are typically obtained from the same sample, from the same study procedures, and/or in a similar context. In other words, these effect sizes are dependent. It is well known that ignoring the correlation between effect sizes may lead to flawed inferences due to underestimation of standard errors, which in turn leads to an increase of the likelihood of false positives (Becker, 2000; Hedges, 2009). To appropriately account for dependency among effect sizes within studies, traditional two-level models can be extended to three-level models (Cheung, 2014; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013, 2015), adding an intermediate level that explicitly accounts for the variation among effect sizes within studies:

$$d_{jk} = \delta_{00} + r_{jk} + w_{jk} + u_k \tag{4}$$

d_{jk} refers to the effect size of outcome j in study k. Whereas in the traditional random effects model (Eq. 3) there is only one random effect whose variance has to be estimated (i.e., between-studies variance), in this extended model there are two random effects, both of which are assumed to follow a normal distribution with zero mean. The first random effect, w_{jk}, refers to the deviation of the population effect j in study k from the mean population effect in study k, whereas the second random effect, u_k, refers to the deviation of the study mean from the overall mean population effect. Therefore, σ^2_w, represents the variability between population effect sizes studied in the same study. Previous papers have referred to this variance as the between-outcomes or within-study variance. The variance σ^2_u refers to the between-studies variance or, in other words, to the variability of study means around the grand population mean (level 3). An advantage of this model is that it assumes that each observed effect size estimates its own population value, so the types of outcomes reported can greatly differ across and within primary studies. Furthermore, unlike in other approaches (e.g., the multivariate approach of Kalaian & Raudenbush, 1996), with a three-level model it is not necessary to estimate the covariances between effect sizes in advance, which are rarely available. Therefore, the application of multilevel models in meta-analysis is especially advantageous when the outcomes reported are very different from study to study, and when studies do not report enough information to estimate the covariance among effect sizes.

Dependency can also occur across studies: studies carried out by the same research group will probably lead to more similar (therefore dependent) effect sizes, or effect sizes from studies done in the same country will correlate more among

them. Therefore, three-level models can be applied also to account for dependencies across studies (Konstantopoulos, 2011). In fact, more levels can be added to account for additional sources of dependencies. For instance, the two scenarios presented before could be combined: multiple effect sizes (level 2) might be nested within studies (level 3), and studies, at the same time, might be nested within different countries (level 4). Therefore, in this example, a four-level model should be specified instead of a three-level model. In this four-level model, there are three variances to be estimated: between-outcomes variance (level 2), between-studies variance (level 3), and between-countries variance (level 4).

Other multilevel, but not hierarchical, models have been also proposed for meta-analysis, such as cross-classified random effects models (CCREMs; Fernández-Castilla et al., 2018), showing again the flexibility of the multilevel approach. CCREMs are appropriate when effect sizes are nested within two (or more) types of higher-level units (i.e., random factors) that are not nested within each other, but rather crossed. For instance, if studies give multiple effect sizes for multiple countries, studies are not nested within countries (because one study can give effect sizes for multiple countries), nor are countries nested within studies (because one country can be studied in multiple studies). In other words, effect sizes are nested in a combination (a cross-classification) of countries and studies. Although the use of a model with crossed random effects is the most appropriate approach under some scenarios (see Fernández-Castilla et al., 2018 for more examples), it is difficult to find meta-analyses that apply these models.

Although multilevel models are very flexible, we suspect that applied researchers do not take full advantage of their possibilities. Therefore, two goals of this study are to first describe how multilevel models (with more than one random effect[1]) are typically applied in meta-analysis, and then illustrate how, in some meta-analyses, more sophisticated models could have been applied that account better for the (non-)hierarchical data structure.

Besides multilevel modeling, other techniques exist for dealing with multiple effect sizes, such as multivariate methods (Kalaian & Raudenbush, 1996; Raudenbush, Becker, & Kalaian, 1988) and the Robust Variance Estimation method (RVE; Hedges, Tipton, & Johnson, 2010). For a more comprehensive overview of the suitability of these methods in different circumstances, we refer to López-López, Page, Lipsey, and Higgins (2018). Roughly, the multivariate approach can be applied when there is information about the correlation between the different variables included in a primary study, which is unfortunately rarely the case. Furthermore, if there are many different types of outcomes across studies (see, for instance, Geeraert, Van den Noortgate, Grietens, & Onghena, 2004, where there are up to 52 outcomes in one study) the implementation of this method is even more cumbersome, as information on much more correlations need to be available in order to calculate the sampling covariance between all pairs of effect sizes. If there are many different types of outcomes across studies, an alternative is to apply a three- (or more) level model or to apply RVE method, as in these methods it is not necessary to know the correlation between outcome variables. Simulation studies have shown that these methods perform similarly (Moeyaert et al., 2017). The main difference is that RVE is typically used with a standard random effects model (and therefore gives an estimate of the total variance only) but robust standard errors are estimated, whereas in multilevel models additional random effects are included to account for the dependency, resulting in separate variance estimates for each random effect.

Several simulation studies have explored and compared the performance of these methods (e.g., Fernández-Castilla et al., 2018; Hedges et al., 2010; Lee, 2014; Moeyaert et al., 2017; Park & Beretvas, 2018; Tipton, 2013, 2015; Van den Noortgate et al., 2013, 2015). The conditions generated in these simulation studies were to a certain extent based on results of real meta-analyses (see a summary of the selected conditions in Table 1). However, a systematic study of what values can be expected is lacking. Therefore, the third goal of this study is to give a full description of the main characteristics of meta-analyses of multiple outcomes (that use multilevel modeling techniques to carry out the synthesis). This information will be given in function of the research field so that future simulation studies can generate and explore realistic scenarios. The only information available about characteristics of these meta-analyses has been given by Park and Beretvas (2018). However, the main drawback of this review is that the authors extracted the information from meta-analyses published on a specific journal in the field of education, meaning that the search was not systematic and only applies to that field. There is another systematic review that describes characteristics of meta-analyses of clinical psychology treatments (Rubio-Aparicio, Marín-Martínez, Sánchez-Meca, & López-López, 2018), but the authors only selected the most relevant effect size per primary study, meaning that this review does not include information about the distribution of the number of outcomes per study.

Method

Search procedure

Meta-analyses applying a multilevel model with more than one random effect (i.e., three-, four- or five- level models

[1] Multilevel models with more than one random factor refer to three-, four-, and five-level models and to CCREMs. We explicitly state that the scope of this systematic review are meta-analyses that apply 'multilevel models with multiple random effects' because we want to exclude meta-analyses that apply a traditional random effects model, which is a multilevel model with only one random effect.

Table 1 Characteristics of published simulation studies that explore the performance of several methods for dealing with multiple effect sizes

	Effect size	ES value	Studies	Sample size of one group	Number of outcomes	Total variance	Moderators
V.D.N. et al. (2013)	SMD	0, 0.20, 0.40	30, 60	Balanced: 25, 50	Balanced: 2, 5	0.10	No
V.D.N. et al. (2015)	SMD	0.40	30, 60	Balanced: 25, 50	Balanced: 7 Unbalanced: 1-7	0.10	No
Lee (2014)	SMD	0.20, 0.40	30, 60	Balanced: 25, 50	Balanced: 2, 5	0.10	No
Moeyaert et al. (2017)	SMD	0, 0.20, 0.40	30, 50	Balanced: 25, 50	Balanced: 2, 4	0.30	No
Park and Beretvas (2018)	SMD	Three different outcomes: .2, .4 and .6	20, 50	Unbalanced: a) Median=25; b) Median=180	Balanced: 15 Unbalanced= 1-15	Ratio: 5:3:2 and 2:3:5	One moderator
Hedges et al. (2010)	SMD	0.5	10, 20, 40	Balanced: 10, 20, 40	Balanced: 1, 2, 5, 10 Unbalanced: a) 1 or 10; b) 1 or 5	0, 0.5v, or 1v	One moderator
Tipton (2013)	1. Risk difference 2. Log-Odds ratio 3. Log-risk ratio	Three sets of probabilities: (pC, pT) = (0.10, 0.10), (0.50, 0.50), (0.30, 0.40) Varying case: (pC = 0.10, 0.30, and 0.50) and pT = 0.20, 0.40, and 0.55	10, 20, 40	Balanced: 20, 50, 100 Unbalanced: a) 20 or 100, b) 20, 40, 60, 80 or 100.	Balanced: 1, 2, 5, 10. Unbalanced: a) 1, 2, 5 or 10; b) 1 or 10 (1 in a larger %)	$I^2_{studies}$ = 0, 0.33, 0.5	One moderator
Tipton (2015): Study 1	SMD		10, 20, 40	40	Balanced: 1, 2, 5 Unbalanced: within meta-analysis, studies could report 1, 2, 5 or 10 outcomes	$I^2_{studies}$ = 0, 0.33, 0.5	One moderator, unbalanced. It could be continuous or binary
Tipton (2015): Study 2	SMD		20	40 and unbalanced across studies	40 and unbalanced across studies		Four covariates at the same type in the regression

SMD standardized mean difference; *ES* effect size; *V.D.N.* Van den Noortgate

and CCREMs) were searched in July 2018. First, we systematically searched for meta-analyses in six electronic databases: Web of Science, Science Direct, Medline PubMed, Psychology Database, Scopus and ERIC. The search strings used were "three-level meta-analysis" OR "multilevel meta-analysis" OR "multilevel meta-analytic review". Second, we looked at meta-analyses citing studies which introduced the use of multilevel techniques for meta-analysis, namely the studies of Hox and de Leeuw (2003) and Raudenbush and Bryk (1985). Third, we searched for meta-analyses that referenced one of the three studies that specifically explain and illustrate the use of three-level models in meta-analysis for dealing with dependent effect sizes. These are the studies of Cheung (2014), Konstantopoulos (2011), and Van den Noortgate et al. (2013, 2015). The second and third step were carried out using Google Scholar. Afterward, the title, abstract or full text was screened to check whether the meta-analyses met the inclusion criteria.

Inclusion/Exclusion criteria

No date restriction was imposed in our search. A meta-analysis was included in this review if: a) results from group studies were combined; meta-analyses of single-case experimental studies were excluded; b) empirical effect sizes were combined using a model with more than one type of random effect, nested and/or crossed; meta-analyses using a traditional random effects model therefore were excluded, because they include only one type of random effect, c) the study was reported in a journal article, dissertation or conference paper; books, posters, or any other format were excluded, d) the paper was written in English, Spanish, or Dutch.

Data extraction

A MS Excel file was created in which several characteristics of the meta-analyses were coded. First, the name of the first author, the country where his/her research group or institute was located, and the research field based on the categorization given by the Web of Science were coded. If the meta-analysis was categorized in several research domains, then we selected the first category. This classification was sometimes too specific, making it difficult to group studies in broader categories. Therefore, we later used the classification of the Research Foundation of Flanders (FWO), which groups disciplines within five large categories: 1. Behavioral and social sciences, 2. Biological sciences, 3. Cultural sciences, 4. Medical sciences and 5. Science and technology[2]. If the meta-analysis was not found in the Web of Science or if the article was labeled as 'multidisciplinary', the first author classified the paper in one of these five domains according to the content of the article.

Afterward, we coded the effect size synthesized and the model fitted: either a three-level model, four-level model, five-level model, or two- or three-level CCREM. The number of units for each type of random effect was also registered (e.g., the number of countries). We were also interested in the sources of dependency among effect sizes; that is, why effect sizes were dependent within and between studies (i.e., multiple outcomes that measure a common construct, multiple treatment groups, multiple follow-ups, etc.).

Next, common characteristics of meta-analyses were coded. In this block, we first coded the number of independent meta-analyses performed in each study. If more than one independent meta-analysis was performed, then we selected the one that included more studies, so only one meta-analysis per paper was registered. Then, we coded the number of studies analyzed, the total sample size used, the value of the pooled effect size and the variance of each random component, together with the I^2 index and the intraclass correlation coefficient (ICC) of each level (Cheung, 2014). In order to give information about common sample size values and typical number of effect sizes reported in primary studies, we also coded in another MS Excel document the sample size and the number of outcomes reported in each primary study within each meta-analysis. Sometimes, within a primary study, it was not possible to know if the same sample of participants was used for all reported outcomes (related samples), or if different samples were used for different outcomes (independent samples). In those cases, we assumed that these samples were related and computed the median sample size of each primary study.

[2] This classification is available in Dutch in the following link: https://t.co/ZtPWCkNh1y

Another characteristic that was coded was whether a moderator analysis was performed, how many moderators were tested, the strategy used to analyze them (e.g., moderators can be added to the regression model one by one, in blocks, simultaneously, etc.), and whether interaction terms were tested. When important information was not reported (e.g., the model fitted) or if inconsistencies were found (e.g., the number of outcomes that the authors reported did not match the ones reported in the supplement material), the corresponding authors were contacted. In total, we contacted 20 authors, and 16 of them replied.

All studies were coded by the first author; the second author independently coded 20% of the studies. Afterward, interrater agreement was calculated by dividing the number of agreements by the sum of the number of agreements plus the number of disagreements. Disagreements were solved through discussion.

All analyses were done in R, and the datasets are available in https://osf.io/znc68/

Results

Characteristics of meta-analyses with multiple random factors

Table 2 shows how many meta-analyses were first retrieved, how many of them were eliminated after a first screening (and the reasons why), and finally how many were kept after a second, more comprehensive screening of the full text. The main reasons for excluding an article were that it was not a meta-analysis, that it was a methodological paper, that it was not a journal article, dissertation or conference paper, that it was already included in other articles, or that the model fitted was a traditional random effects model. A total number of 178 meta-analyses were finally selected for this review. The list of excluded studies is available upon request from the first author.

General characteristics The first three-level meta-analysis was published in 2002 (i.e., Beretvas, Meyers, & Leite, 2002). As can be seen in Fig. 1, there was a substantial increase in the number of publications of multilevel meta-analyses with more than one random factor in 2016. There is a clear increasing tendency: in the first six months of 2018, the same number of meta-analyses with more than one random effect has been published as in the whole 2017.

The interrater agreement resulting from the coding was 71%. The characteristic with most disagreements was the sample size of each primary study within each meta-analysis. As mentioned before, sometimes it was very difficult to know whether, within studies, the sample sizes reported for each

Table 2 Number of studies excluded in each phase and reason for exclusion

	Reference list						
	Hox and de Leeuw (2003)	Raudenbush and Bryk (1985)	Cheung (2014)	Konstantopoulos (2011)	Van den Noortgate et al. (2013)	Van den Noortgate et al. (2015)	Electronic datasets
Retrieved studies	**85**	**267**	**125**	**102**	**109**	**46**	**552**
Deleted after first screening							
Not a meta-analysis	4	38	3	7	-	-	225
About methodology	12	112	26	22	24	9	55
Already retrieved in other study	5	3	39	19	12	1	62
Appeared twice in the same search	-	1	4	1	3	2	101
Not an article	5	51	3	7	5	2	4
Other language	3	5	6	3	2	-	-
Not reachable	-	9	2	1	2	-	4
SCED meta-analysis	-	-	-	-	-	-	10
Duplicated data	-	-	2	-	-	-	-
Retracted meta-analysis	-	-	-	1	-	-	-
Screened studies	**56**	**48**	**40**	**41**	**61**	**32**	**91**
Deleted after full screening							
Use a traditional random effects model	40	41	10	12	7	7	58
Apply other technique (RVE, multivariate)	-	1	2	2	4	-	5
Duplicated data	4	1	-	-	-	-	-
Combine raw data, no studies	-	-	-	-	-	-	1
Final number of studies	**15**	**5**	**28**	**27**	**50**	**25**	**28**

SCED Single-case experimental design; *RVE* robust variance estimation method. "Electronic datasets" refers to the articles found in Web of Science, Science Direct, Medline PubMed, Psychology Database, Scopus, and ERIC

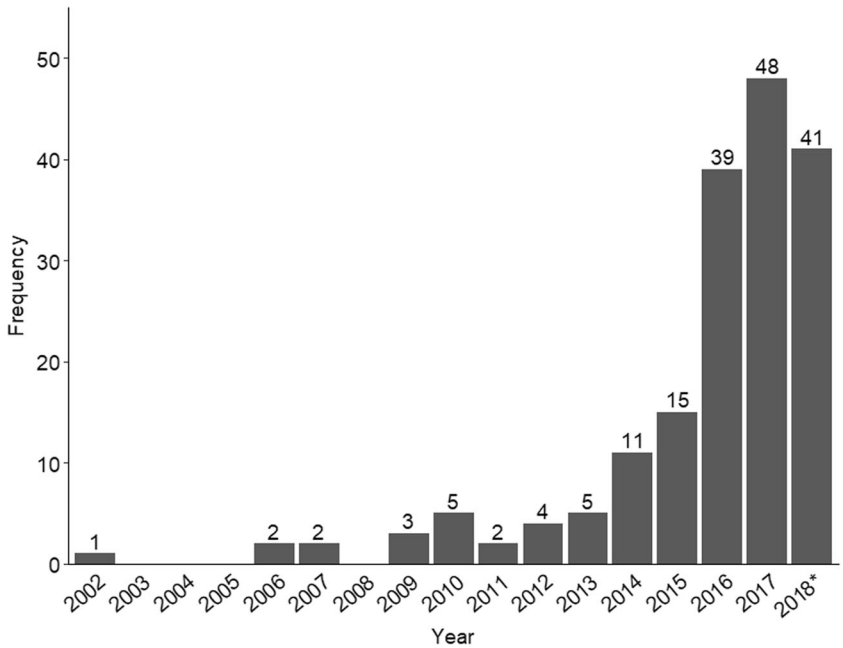

Fig. 1 Number of published meta-analyses with multiple random effects by year. Notice that the search was done in July of 2018, so this bar only includes data from January to July of 2018

outcome were independent or related. All disagreements were discussed until the two judges agreed.

From the 178 meta-analyses retrieved, 121 studies belonged to the behavioral and social sciences field (e.g., psychology, economics, law), 33 studies to biological sciences (e.g., biology, ecology, nutrition), 20 studies to medical sciences (e.g., medicine, biomedicine), two meta-analyses to cultural sciences (e.g., languages, history), and two studies to science and technology (e.g., computer sciences). Due to the small sample of studies belonging to these last disciplines, they will not be discussed in the Results section.

Regarding the software used, most meta-analyses used R software. Specifically, 82 studies used the package *metafor*, 28 used *metaSEM*, two performed the meta-analysis with the *MCMCglmm* package, one used the *brms* package, one used the *nlme* package, and one used the *rStan* package. Following the R software as the most used software are SAS (25), HLM (7), Stata (4), S-Plus (4), MLwinN (2), WinBUGS (1) and M-plus (1). There were 19 meta-analyses that did not mention the software used. The model most commonly fitted was a three-level model: 162 meta-analyses used a three-level model, eight meta-analyses fitted a four-level model, six studies fitted a CCREM, and two studies used five-level models.

Figure 2 shows how many times each type of effect size was used to synthesize study results. As shown, the Pearson correlation coefficient was the effect size most commonly combined, followed by Cohen's *d* / Hedges' *g* (for independent samples), Cohen's *d* for related samples, Fisher's *z* and odds ratios. A variety of other effect sizes or summary statistics (e.g., percentages, standardized means, rates, or R^2) were less commonly used. In order to simplify the following analyses, we have put Hedges' *g* and Cohen's *d* (pre-post) in the same category as Cohen's *d*. The eight Fisher's *z*-values were back-transformed to Pearson correlation coefficients, and were discussed together with these correlation coefficients.

Number of studies In the field of behavioral and social sciences and in biological sciences, the range of number of primary studies included in the meta-analyses was quite wide (see Table 3). Furthermore, these were the fields in which more primary studies were included on average in the meta-analyses, and the distribution of these studies was positively skewed with a heavy tail. In contrast, in the field of medical sciences the number of primary studies analyzed varied less and was on average smaller. In these cases, the median and the mean were very similar, indicating that the distribution was closer to a normal distribution.

Sample size of primary studies Regarding the distribution of the sample sizes of the primary studies within each meta-analysis (Table 3), there was one (outlying) primary study in behavioral and social sciences that included 1,957,491 observations. Also, there was a large discrepancy between the mean and the median: the median was much smaller than the mean in all disciplines, indicating that the distribution of sample sizes of primary studies was very positively skewed, or in other words that small sample sizes were more common than large sample sizes. In medical science and in behavioral and social sciences, the sample sizes were larger than in biological sciences.

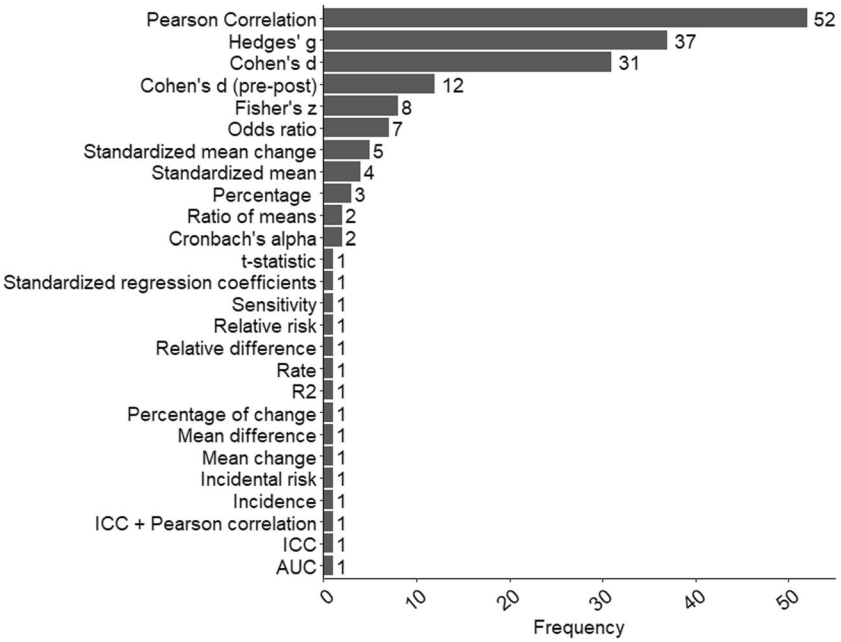

Fig. 2 Effect sizes and summary statistics synthesized in each meta-analysis

Table 3 Characteristics of the 178 meta-analyses included in this systematic review

		Min	1st Qu.	Median	Mean	3rd Qu.	Max
Number of studies	*m*						
Beha. & Soc. Sci.	116	5	23	39	64.69	69	456
Biological Sci.	32	8	16	23	43.38	43	298
Medical Sci.	20	6	15	32	35.40	48	88
Sample size of primary studies	*k*						
Beha. & Soc. Sci.	3471	5	50	107	2,003.15	284	1,957,491
Biological Sci.	710	2	12	25	96.82	56	5,320
Medical Sci.	246	6	26	124	2,167.77	601	128,681
Outcomes in primary studies	*k*						
Beha. & Soc. Sci.	3818	1	1	2	3.56	4	76
Biological Sci.	860	1	1	2	6.22	5	202
Medical Sci.	369	1	1	1	2.40	2	21
Number of moderators	*m*						
Beha. & Soc. Sci.	121	0	5	8	8.98	13	31
Biological Sci.	33	0	4	6	6.97	8	33
Medical Sci.	20	0	3	6	6.75	8	17

m meta-analyses; *k* primary studies

Number of effect sizes Regarding the distribution of outcomes within studies (Table 3), we can see that the median number of outcomes was quite small and ranged between 1 and 3 across research disciplines. In the field of biological sciences and behavioral and social sciences, the maximum number of outcomes reported in a primary study reached very high values compared to maximum number of outcomes observed in the medical sciences. Figure 3 shows the percentage of primary studies in each field which included 1, 2, 3, 4, or 5 or more outcomes per study.

Types of dependencies modeled From the 178 papers, there were 115 meta-analyses (64.61%) reporting that (only) one source of dependency between effect sizes existed. In 49 meta-analyses (27.53%), authors reported two reasons why effect sizes were dependent, and in 13 (7.30%) and in one (0.56%) meta-analyses authors said that there were three and four sources of dependency, respectively.

The type of dependency most commonly reported was the existence of multiple effect sizes related to a common construct (127 meta-analyses). In 32 meta-analyses, several effect sizes were reported within primary studies because several treatment groups were compared to a common control group. In 25 papers, several sub-samples were used within primary studies. Another source of dependency that was reported in 22 studies was the existence of multiple follow-ups, that is, two

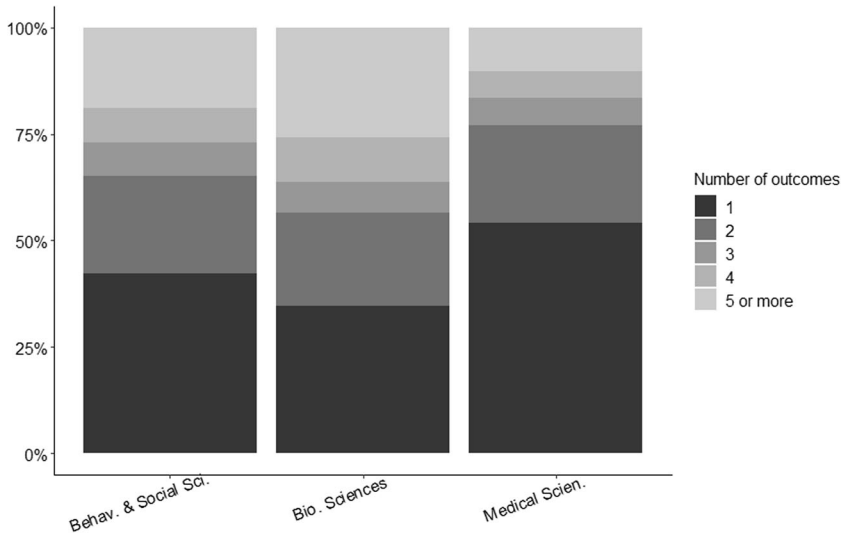

Fig. 3 Percentage of primary studies that include 1, 2, 3, 4 or 5 or more effect sizes by discipline

groups were repeatedly compared at different time points. In 13 meta-analyses, authors reported that multiple effect sizes were extracted because primary studies used several instruments to measure a common construct. Finally, the authors of 12 meta-analyses said that due to the existence of several experiments and conditions within primary studies, more than one effect size could be retrieved.

In meta-analyses where dependency occurred across studies, the most common type of dependency was that papers were nested within countries (eight meta-analyses). In six papers, studies were dependent because they belonged to the same author or were performed by the same research group, and in ten other meta-analyses, studies were grouped in different higher-level factors, such as phylogenetic families, years in which studies were performed, etc.

Pooled effect distribution Table 4 shows descriptive information of the overall effect size values in function of the type of effect size and scientific discipline. Looking at the pooled Cohen's *d*, we can see that the mean and median were really similar in all disciplines, indicating that the distribution was close to being normal. Most available data came from behavioral and social sciences, where in general effect sizes were quite small if we compare them to the cutoffs proposed by Cohen (1988). In the field of biological sciences, the values were slightly higher, but still small if we take as a reference Cohen's values. In medical sciences, the range of possible values that Cohen's *d* could take was smaller compared to the other disciplines.

Table 4 Descriptive of each type of effect sizes in function of the discipline

	m	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Cohen's *d*							
Beh. & Soc. Sci.	55	– 0.67	0.10	0.25	0.21	0.37	1.02
Bio. Sci.	12	– 1.06	0.18	0.35	0.41	0.52	1.80
Medical Sci.	8	– 0.35	– 0.11	0.24	0.20	0.48	0.67
Pearson corr.							
Beh. & Soc. Sci.	42	– 0.21	0.01	0.17	0.20	0.33	0.84
Bio. Sci.	6	– 0.02	0.19	0.24	0.27	0.28	0.72
Medical Sci.	3	– 0.32	– 0.22	– 0.11	– 0.12	– 0.03	0.06
Odds ratio							
Beh. & Soc. Sci.	2	1.19	1.20	1.20	1.20	1.21	1.21
Bio. Sci.	1	2.02	2.02	2.02	2.02	2.02	2.02
Medical Sci.	3	– 1.56	– 0.27	1.02	0.60	1.68	2.34
SMC							
Beh. & Soc. Sci.	3	– 0.04	– 0.02	0.01	0.04	0.09	0.16
Bio. Sci.	0	-	-	-	-	-	-
Medical Sci.	2	– 0.07	0.00	0.08	0.08	0.15	0.23

m meta-analyses; *SMC* standardized mean change

Looking at Pearson correlations, we observed that the means and the medians are similar. In the three disciplines, the median pooled correlations could be considered as medium to small (using the rules of thumb of Cohen, 1988). For the studies using the odds ratio, little information was available. As shown in Table 4, the pooled odds ratio was larger in biological sciences, and in medical sciences there was a higher variability whereas in biological sciences and behavioral and social sciences the combined odds ratio was in general homogeneous among meta-analyses. Finally, there was little information available regarding the standardized mean changes and only for the research disciplines of behavioral and social sciences and medical sciences. The median pooled standardized mean difference was in general very small and homogeneous.

Heterogeneity of effects Table 5 shows the level 2 and level 3 variance estimates and the ICCs of the meta-analyses that fitted a three-level model. The ICCs are calculated following the indications of Cheung (2014) and using the median sampling variance. We have not included standardized mean changes in these results because we did not have enough data. A common result across disciplines and types of effect sizes was that the median variance was, in general, smaller than the mean, indicating that the distributions of the variances at all levels were positively skewed.

Regarding Cohen's *d*, variances were larger in biological sciences and in the medical sciences field compared to behavioral and social sciences. However, the maximum variances were observed in behavioral social sciences, indicating that the skewed distribution of the variances had a heavier tail in this field. For behavioral and social sciences and medicine, we could calculate the median I^2 index using the median sampling variance, the second level variance and the third level variance. The resulting $I^2_{(2)}$ and $I^2_{(3)}$ indexes for behavioral and social sciences were 17.0% and 60.9% respectively, representing the proportion of the total variation of the effect sizes due to second and third level heterogeneity. Therefore, the ratio of variances (median sampling variance : level 2 variance : level 3 variance) was 2:2:6 approximately. In medical sciences, both I^2 indexes equaled 45.8%, and therefore the ratio of variances was 1:4:5 approximately.

Regarding Pearson correlations, in the three disciplines the between-studies variances were larger than the within-study variances. In behavioral and social sciences and in medical sciences, the range of possible values that the three variances could take was quite narrow, while in biological sciences the range of values was wider for the level 3 variance. We could calculate the I^2 indexes for the behavioral and social sciences field: $I^2_{(2)}$ was equal to 32.3%, and $I^2_{(3)}$ was 50. Therefore, the ratio of variances was 2:3:5.

Table 5 Descriptive information of variance components estimates for three-level models

	m	Min	1st Q.	Median	Mean	3rd Q.	Max
Cohen's *d*							
Behav. & Social Sciences							
Median samp. var.	15	0.001	0.016	0.026	0.071	0.048	0.571
level 2 variance	39	0.000	0.009	0.020	0.145	0.088	1.630
level 3 variance	40	0.000	0.017	0.072	0.118	0.153	0.553
$ICC_{(2)}$ / $ICC_{(3)}$				21.74% /78.26%			
Biological sciences							
Median samp. var.	0	-	-	-	-	-	-
level 2 variance	2	0.000	0.019	0.039	0.039	0.058	0.077
level 3 variance	2	0.151	0.206	0.261	0.261	0.315	0.370
$ICC_{(2)}$ / $ICC_{(3)}$				13.00%/87.00%			
Medical sciences							
Median samp. var.	1	0.022	0.022	0.022	0.022	0.022	0.022
level 2 variance	3	0.094	0.107	0.120	0.211	0.269	0.418
level 3 variance	3	0.031	0.076	0.120	0.109	0.148	0.176
$ICC_{(2)}$ / $ICC_{(3)}$				50.00%/50.00%			
Pearson correlation							
Behavioral & social sciences							
Median samp. var.	14	0.000	0.002	0.006	0.008	0.011	0.030
level 2 variance	29	0.000	0.007	0.011	0.018	0.029	0.080
level 3 variance	30	0.000	0.008	0.017	0.022	0.029	0.080
$ICC_{(2)}$ / $ICC_{(3)}$				39.29%/60.71%			
Biological sciences							
Median samp. var.	0	-	-	-	-	-	-
level 2 variance	3	0.000	0.006	0.011	0.010	0.015	0.019
level 3 variance	3	0.029	0.105	0.180	0.216	0.310	0.440
$ICC_{(2)}$ / $ICC_{(3)}$				5.76%/94.24%			
Medical sciences							
Median samp. var.	0	-	-	-	-	-	-
level 2 variance	2	0.010	0.021	0.033	0.033	0.044	0.055
level 3 variance	2	0.040	0.046	0.053	0.053	0.059	0.065
$ICC_{(2)}$ / $ICC_{(3)}$				38.37%/61.63%			
Odds ratio							
Behav. & Social Sciences							
Median samp. var.	1	0.030	0.030	0.030	0.030	0.030	0.030
level 2 variance	2	0.028	0.029	0.029	0.029	0.030	0.030
level 3 variance	2	0.040	0.047	0.054	0.054	0.060	0.067
$ICC_{(2)}$ / $ICC_{(3)}$				34.94%/65.06%			
Medical Sciences							
Median samp. var.	0	-	-	-	-	-	-
level 2 variance	1	0.060	0.060	0.060	0.060	0.060	0.060
level 3 variance	1	0.190	0.190	0.190	0.190	0.190	0.190
$ICC_{(2)}$ / $ICC_{(3)}$				24.00%/76.00%			

m number of meta-analyses; *median samp. var.* median sampling variance. $ICC_{(2)}$ intraclass correlation coefficient at level 2. Proportion of the total variance due to level 2 units. $ICC_{(3)}$ intraclass correlation coefficient at level 3. Proportion of the total variance due to level 3 units

Table 6 Descriptive information of variance components estimates for four-level models

	m	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Behavioral & social sciences							
Cohen's d							
Median samp. var.	1	0.032	0.032	0.032	0.032	0.032	0.032
level 2 variance	2	0.020	0.027	0.034	0.034	0.041	0.048
level 3 variance	2	0.015	0.029	0.043	0.043	0.056	0.070
level 4 variance	2	0.054	0.066	0.077	0.077	0.089	0.100
Pearson correlation							
Median samp. var.	0	-	-	-	-	-	-
level 2 variance	1	0.008	0.008	0.008	0.008	0.008	0.008
level 3 variance	1	0.015	0.015	0.015	0.015	0.015	0.015
level 4 variance	1	0.000	0.000	0.000	0.000	0.000	0.000

m number of meta-analyses; *median samp. var.* median sampling variance

Looking at odds ratio, we can see that the variance values in medical sciences were relatively larger than the variance values in behavioral and social sciences. The $I^2_{(2)}$ in behavioral and social sciences was 25.8%, and $I^2_{(3)} = 47.6\%$. Therefore, the ratio of variances was 2:3:5. From the eight studies that fitted a four-level model, only three studies from the behavioral and social sciences field reported values for variance components estimates. Two studies report Cohen's *d,* and one Pearson correlations. Information about the values of these variances is reported in Table 6.

Finally, from the six studies that reported CCREMs, only two reported the variance component estimates. These two meta-analyses belonged to the discipline of science and technology, reporting a Pearson correlation, and biological sciences, reporting a Cohen's *d*. Both fitted a three-level model, with a crossed-classification at third level. In the study that belonged to science and technology, the second level variance equaled 0.12, and the variance of the crossed-factors at third level were 0.14 and 0.19. Regarding the meta-analysis of biological sciences, the second level variance was 0.13, and the variance of the crossed-random factors at third level were 0.07 and 0.10.

Moderators and strategies Table 3 shows the distribution of the number of moderators analyzed in each discipline. On average, in behavioral and social sciences, the effect of eight moderators was typically tested, whereas in other disciplines the number was slightly smaller, between five and six. The distribution of moderators was especially positively skewed in the disciplines of behavioral and social sciences and in biological sciences, where the maximum number of moderators tested was around 30. In medical sciences, the maximum number of moderators was much smaller. Also, in 32 meta-analyses, interactions among moderators were tested in the multilevel model.

Most studies (41.01%) did separate analyses for each moderator. The second most popular strategy for analyzing moderator effects was to introduce all moderators simultaneously (15.16%). Another strategy was to introduce the moderators one by one, and then fit a final meta-regression with all moderators that resulted to be significant (11.23%). The fourth most common strategy was to introduce separate blocks of moderators (7.86%). There was a relationship between these four strategies and the number of moderators tested in the meta-analysis ($F = 3.42$, $p < .05$). Specifically, the strategy consisting in introducing moderators one by one in the model and then test all significant moderators simultaneously, is more often used when many moderators were tested (mean number of moderators = 11.80, standard deviation= 7.95) compared to when the strategy was to analyze them simultaneously (mean = 6.70, standard deviation = 5.15).

Other strategies that were used in a lesser extent were: introducing moderators one by one in the regression, and then in a second step introduce all moderators simultaneously, including the ones that were not significant in the first step (2.80%), introducing all moderators simultaneously, and then removing the ones that were not significant (2.80%), and introducing moderators one by one and then in separate blocks (1.12%). Only one meta-analysis used the MetaForest technique (Van Lissa, 2017), which consists in the application of the random forest technique in meta-analysis, and one meta-analysis (Hijbeek, van Ittersum, ten Berge, Gort, Spiegel, & Whitmore, 2017) used a technique called multi-model dredging.

Description of the use of multilevel models in meta-analysis

The model most commonly fitted was the three-level model: 162 meta-analyses applied a three-level model, eight meta-analyses fitted a four-level model, six studies fitted a

CCREM, and two studies reported five-level models. Within the six meta-analyses that applied a CCREMs, two of them applied a two-level CCREM. Two other studies applied a three-level CCREM where random factors were crossed at third level. Another meta-analysis fitted a four-level model with a crossed-classification at the second level. Finally, in the study of Golivets and Wallin (2018) there were up to eight random components included in the model, without being completely clear how many levels there were. At least four of these random effects were crossed at level 2.

From the meta-analyses applying a three-level model, 123 (69.10%) selected as the third level factor 'studies', that is, these meta-analyses modeled the dependency of effect sizes that belong to the same study. Fifteen meta-analyses (8.43%) and six meta-analyses (3.37%) modeled the dependency among effect sizes within 'samples' and within 'countries', respectively. Other factors at third level were: authors, datasets, experiments, laboratories, societies, species, etc.

Four-level models were typically applied to treat additional sources of dependencies within outcomes, such as dependency among comparisons between several treatment groups (e.g., Krieger, 2010; Lehtonen, Soveri, Laine, Järvenpää, de Bruin & Antfolk, 2018; Schoenfeld, Ogborn, & Krieger, 2015) or dependency of effect sizes across sub-scores and tasks (Fradkin, Strauss, Pereg, & Huppert, 2018), and to treat dependency across studies. For instance, studies could be grouped within higher-level clusters such as culture (e.g., Kende, Phalet, Van den Noortgate, Kara, & Fischer, 2017), laboratory (e.g., Martineau, Ouellet, Kebreab, & Lapierre, 2016) or sub-region (e.g., De Noordhout et al., 2014). In the meta-analysis of Kende et al. (2017), where a four-level model was fitted, we find another example of how flexible multilevel models can be in their specification. Instead of estimating a total between-studies variance at level 3, the authors decided to model separately the between-studies variance (or within-culture variance) of studies carried out in the United States, and the between-studies variance of studies carried out outside the United States, because they expected a greater variation across studies done in the United States. In fact, they found that variance within-United States (0.018) and within-other countries (0.011) was slightly different.

Two meta-analyses from Scharfen, Blum, and Holling (2018) and Scharfen, Peters, and Holling (2018) accounted for five sources of dependency, leading to the application of a five-level model: effect sizes (level 1) were nested within multiple comparisons (level 2), because primary studies could report the comparison of a group across different time points. Comparisons were then nested within different outcomes (level 3), that were nested within sub-samples (level 4) that, finally, were nested within studies (level 5).

There are several examples of the application of CCREMs. For instance, O'Shea and Dickens (2016) fitted a four-level model with a crossed-classification at level 2: effect sizes were nested within outcomes and within instruments. The same outcome could be measured with several instruments, and the same instrument was used to measure different outcomes. Hence, these two random factors were cross-classified. Outcomes and instruments, in turn, were nested within studies (level 3), meaning that they assumed that the types of outcomes and instruments differed across studies. Finally, studies were nested within authors (level 4).

Another interesting example is the meta-analysis of Francis (2016), where a two-level CCREM was fitted with a cross-classification at the second level. In this case, there were several effect sizes within studies, that were obtained using different instruments. In other words, within studies effect sizes could be obtained from several instruments, and at the same time some instruments were used across studies. Because across studies some effect sizes were obtained from the same instrument, these two factors were not purely nested. In fact, effect sizes were nested within studies, and at the same time effect sizes were nested within instruments, meaning that 'instruments' and 'studies' constituted two crossed factors at level 2.

Finally, there is a special case in the meta-analysis of Golivets and Wallin (2018), who specified up to eight random components to account for several dependencies in their data. Most of these random effects were crossed. This meta-analysis studied the competition between plant species. Therefore, each effect size referred to two types of plant: the target plant, and the neighbor plant. The variability of the effect sizes due to the existence of different target species and due to the existence of different neighbor species was modeled with two crossed factors. In addition, species that belonged to the same phylogenetic family were supposed to be more alike, so other two random factors were added for modeling the phylogeny of the target and of the neighbor species. This is an example of how sophisticated a model can become in function of the different sources of dependencies.

Discussion

Considerations for specifying random effects

In the Introduction, we mentioned the advantages of using multilevel models to deal with dependent effect sizes in meta-analyses: the correlation between effect sizes from the same study is not needed (unlike the multivariate approach), separate estimates of the different variance components (e.g., between-studies and within studies) are estimated (unlike in the RVE method), and moderator effects can be allowed to vary across studies (or across any other higher-level cluster).

However, when using multilevel models to deal with dependent effect sizes, it is very important to correctly specify the model according to the (non-) hierarchical structure of the

data to get appropriate parameter estimates. In other words, all relevant random effects must be included in the model to avoid biased estimates (McNeish, Stapleton, & Silverman, 2017), and sometimes it is difficult of decide whether the (moderator) effect of a variable should be considered as fixed or as random. Snijders and Bosker (2012) give some guidelines to take such a decision. A variable can be considered as random if its categories can be seen as a random sample of a population of (interchangeable) units. Moreover, these authors mention that, as a rule of thumb, at least 20 categories are necessary to properly estimate the variance of a random effect. As an example, let us imagine that studies are nested within 20 different countries. If the meta-analyst is not interested in the separate effect for each of these countries but only wants to estimate the extent to which effect sizes vary due to country differences, the country effect can be specified as random. An optional step here would be to try to explain this variance (i.e., the variability of effect sizes due to differences between countries) by including variables with a fixed effect (e.g., a variable that indicates whether studies are from United States or from Europe). Sometimes, however, a variable does not have enough categories to consider it as a random effect (e.g., when studies are nested within countries, but there are only five different countries), or researchers might be interested in the separate overall effect size for each country. In this scenario, the variable "country" can be introduced in the model as a moderator with a fixed effect (i.e., estimating one separate effect for each of the five countries or taking one country as a reference and estimating the contrast of the effect in this country with the ones in the other countries).

As mentioned before, a few studies found that the RVE method performs similarly as multilevel models. An advantage of applying RVE is that it only requires the correct specification of the higher-level cluster (McNeish et al., 2017). Therefore, this method is a good alternative when the researcher is unsure of the correct model specification. However, this method only gives an estimate of the total variance, and not separate estimates for each random effect. An interesting approach was recently proposed by Tipton, Pustejovsky, and Ahmadi (2019): applying first a multilevel model, and therefore getting separate estimates for the variance components, and then applying RVE method to get robust standard errors. Although this approach seems promising, simulation studies still have to investigate the performance of this approach.

Examples of alternative specification of multilevel models in meta-analysis

In this section, we identify five common situations in which models with more random effects than the basic hierarchical three-level model would have been more appropriate given the (non-)hierarchical data structure. We only give some prototypical example for illustrative proposes, but there are more meta-analyses included in the systematic review that fit in one of the following categories. First, a fourth level could have been added to model dependency within studies. For instance, Acar and Sen (2013), that studied the relationship between creativity and schizotypy, specified a three-level model where effect sizes (level 1: 268 effect sizes) were nested within studies (level 2: 45 studies), and studies were nested within authors (level 3: 34 authors). This model ignores that within studies, effect sizes might not only vary due to the known sampling variance, but also because they represent different population outcomes. The omission of these between-outcomes (or within-study) variability could lead to biased standard error estimates of the combined effect and of the moderator variables that referred to the outcome variables (Van den Noortgate, Opdenakker, & Onghena, 2005). Adding an additional second level that modeled the between-outcomes variance (or the between-population effect sizes variance) could have been statistically better: effect sizes (level 1, sampling level: 268 effect sizes) are nested within outcomes or within population effect sizes (level 2: 268 outcomes), outcomes (or population effect sizes) are nested within studies (level 3: 45 studies) and studies are nested within authors (level 4: 34 authors).

Second, within studies, more levels could have been specified to deal with different types of dependencies**.** The meta-analysis of Soveri, Antfolk, Karlsson, Salo, and Laine (2017), tests the efficacy of working memory training. Looking specifically at the analyses done on the outcome 'fluid intelligence', we see that authors fitted a meta-analytic three-level model because observed effect sizes (level 1, sampling level: 133 effect sizes) referred to specific population effect sizes or outcomes (level 2: 133 outcomes), that were at the same time nested within studies (level 3: 25 studies). However, primary studies sometimes used more than one treatment group, and the comparison of these treatment groups with a common control group on a specific outcome led to multiple effect sizes. Authors decided to perform a fixed-effect meta-analysis on these multiple effect sizes within comparisons with the aim of having just one single effect size per outcome. This strategy of summarizing effect sizes within a higher-level unit was already proposed by Cooper (2015), called 'strategy of shifting unit of analysis'[3]. The main drawbacks of this approach are that 1) it involves a loss of information and a reduction of power; 2) moderators that refer to the effect sizes within studies cannot be included, and 3) simulation studies have shown that the between-studies variance estimate is artificially reduced when this strategy is implemented (i.e., Moeyaert et al., 2017). A valid alternative is to add a level in

[3] These authors acknowledge that they could make the multilevel model more sophisticated, but decided to reduce it to a three-level model in the name of parsimony

the multilevel model that accounts for dependency between effect sizes due to several comparisons: effect sizes (level 1 – sampling level) are nested within outcomes or within population effect sizes (level 2) that are nested within comparisons (level 3) that are nested within studies (level 4). This approach allows the use of all effect sizes and the incorporation of moderator's variables that refer to characteristics of the effect sizes within studies.

Third, a fourth level could have been added to take into account dependency across samples. Some meta-analyses considered that effect sizes belonging to different samples from the same study were independent. Lebuda, Zabelina, and Karwowski (2016) explore the link between mindfulness and creativity. In their meta-analysis, 89 effect sizes were nested within 20 samples. These samples were nested in 13 studies, but the variability of samples within studies was not modeled. Lebuda et al. (2016) made their dataset available, so we were able to re-analyze the data and specify a four-level model instead of a three-level model. When a three-level model was fitted (ignoring that samples were nested within studies), the combined effect size was 0.219, with a standard error of 0.065, and the between-outcomes variance equaled 0.029 and the between-samples variance was 0.066. When a four-level model was fitted, the pooled effect became a little bit larger (0.239) and the standard error slightly increased (0.070). The between-studies variance was 0.014, and the between-sample variance decreased (0.054) while the between-outcomes variance remained equal (0.029). Although the conclusions of the meta-analysis did not change, we can clearly see how the standard error was somehow shrunken due to the omission of the upper study-level. Also, it is important to note that 13 studies might not be enough to properly estimate the between-studies variance.

Fourth, a five-level model could have been applied to model additional within-study and/or between-study dependencies. The meta-analysis of Rabl, Jayasinghe, Gerhart, and Kühlmann (2014), explores country differences in the relationship between high-performance work system and business performance. A three-level model was fitted, where several effect sizes were nested within 156 studies (level 2), nested within 30 countries (level 3). There were several effect sizes within studies, and therefore the authors decided to calculate a linear composite correlation of these within-study effect sizes to avoid dependency. Another option would have been to add an additional level that modeled the variance between these correlations, and in this way preserve all data. Furthermore, some studies used the same dataset, and the authors averaged the effect sizes of studies that used the same dataset. Instead of averaging these effects across studies, an additional, fifth level could have been added that accounted for the between-datasets variance. In summary, the following five-level model could have been fitted: effect sizes (level 1) nested within outcomes (level 2), nested within studies (level 3), nested within datasets (level 4), nested within countries (level 5).

Fifth and last, CCREM's could have been applied instead of three-level models. Pearce (2017) explored whether exposure to scary TV was related to internalizing behaviors in children (e.g., anxiety, stress, depression). Within primary studies, several effect sizes referring to different internalizing behaviors were reported, and also several effect sizes that referred to the same behavior but were measured with different instruments. A three-level model was fitted to account for these dependent outcomes. The author wanted to control for the fact that different instruments were used to measure the same behavior within studies (pg. 70). However, there were too many different instruments to consider this variable as a moderator (the author mentions nine). Furthermore, some instruments were used for only one effect size, making it again difficult to use this variable as a moderator. Therefore, the variable "scale" was not used in the analyses. An alternative would have been to consider "scales" as a random effect, crossed with studies: effect sizes (level 1) are nested within outcomes, and then outcomes are nested, at the same time, within studies and instruments (level 3). Instruments were not nested within studies because they could have been used in several studies. Also, studies are not nested within instruments because within one study, several instruments were used. The description of a similar example can be found in Fernández-Castilla et al. (2018). Considering "scales" as a random factor would have allowed the researcher to have a measure of how effect sizes vary due to the use of different instruments.

Final discussion

This study describes for the first time how multilevel models are applied in meta-analysis and what their most common characteristics are in function of the research discipline. The first conclusion of this systematic review is that three-level models are often systematically applied although other more complex and sophisticated models are sometimes more appropriate given the meta-analytic data structure. The most likely reason for this is that the methodological papers that explain and illustrate the use and application of multilevel methods for meta-analysis focus only on the three-level model (e.g., Assink & Wibbelink, 2016; Cheung, 2014; Konstantopoulos, 2011; Van den Noortgate et al., 2013, 2015). In this article, we have given several examples of how some meta-analytic data actually have a four- or five-level structure, or even cross-classified. We recommend meta-analysts to carefully study and explore the structure of their data in order to apply a proper model. Furthermore, all these more sophisticated models can be easily fitted with the most commonly used package for performing meta-analysis, namely *metafor* in R (Viechtbauer, 2010). An important warning here is that not all variables can be considered as random

factors. For instance, in one of the examples above, the type of sample within studies was considered as a random effect. We assume that there were many different subsamples (i.e., children, adults, students, workers, etc.) and that the subsamples greatly differed across studies. However, if there were not many different subsamples and the subsamples reported within studies were almost always the same (e.g., men and women), 'subsamples' should not have been considered a random factor, but a fixed factor.

The second aim of this study was to describe the main characteristics of these meta-analyses with multiple random factors. In this systematic review, we have disaggregated this information by research discipline. Results show that in the field of behavioral and social sciences, meta-analyses normally include more primary studies (compared to the other disciples), these primary studies report larger sample sizes, the number of outcomes are more unbalanced across studies, and more moderator variables are tested. Also, the values for the combined Cohen's *d* are, on average, smaller than the cutoffs originally proposed by Cohen (1988). In the medical sciences field, fewer studies are normally included in the meta-analysis, and the number of outcomes reported in primary studies are more balanced. The variability within and between studies is slightly larger than in the field of behavioral and social sciences, although in behavioral and social sciences the distribution of the variance components is more skewed. Meta-analyses in the field of biological sciences are similar to the ones in behavioral and social sciences except in the sample size of primary studies, that is on average smaller. One limitation of this systematic review is that we found only a few studies from the field of cultural sciences and science and technology, so the results for these two fields should be interpreted with caution.

Simulation studies that explore the performance of methods that deal with dependent effect sizes should acknowledge these differences in the characteristics of meta-analyses across disciplines. In fact, another goal of this systematic review was to check whether the simulation factor conditions selected in these methodological papers actually represent characteristics of published meta-analyses. Regarding the effect size value, most simulation studies done on standardized mean differences have selected values of 0, 0.20, and 0.40–0.50. According to this systematic review, these values are fairly close to the 1st quartile, median and 3rd quartile, respectively, of all disciplines except for cultural sciences. Looking at the number of studies, simulation studies have typically generated 10, 20, 30, 40, 50, and 60 primary studies. However, the minimum number of primary studies in behavioral and social sciences, biological sciences and medical sciences is below 10, and this situation has not been contemplated in simulation studies. Even so, the median number of studies included in meta-analysis range between 20 and 39 across disciplines and these values are indeed represented in the simulations. Future simulations should focus on factor conditions where the number of studies is even smaller than ten. This is especially important when multilevel models are applied, because previous research has pointed out that when the number of units at the highest level, typically 'studies', is below ten, multilevel models can lead to inflated Type I error rates (Van den Noortgate & Onghena, 2003).

The number of outcomes reported within primary studies is commonly unbalanced, especially in the field of behavioral and social sciences and in biological sciences. Some simulation studies have only generated balanced data (e.g., Lee, 2014; Moeyaert et al., 2017; Van den Noortgate et al., 2013) which is a very unrealistic scenario. Other simulation studies have generated unbalanced data (e.g., Park & Beretvas, 2018; Van den Noortgate et al., 2015) but in a way in which, from a range of possible values for the number of outcomes, all these values were equally likely (for instance, from a range from 1 to 5, primary studies were equally likely to report 1, 2, 3, 4, or 5 outcomes). However, this systematic review has shown that the largest percentage of primary studies report only one outcome (see Fig. 3). This simulated factor condition is only generated in the simulation study of Tipton (2013), where there was one condition where most of studies included only one outcome, and then some studies reported ten outcomes. This is a realistic condition, especially in the field of behavioral and social science and biological sciences. in medical sciences, the number of outcomes in primary studies is more balanced, as the majority of primary studies only include up to four outcomes.

The sample size of the primary studies has shown to have little impact on the parameter estimates (e.g., Hedges et al., 2010). Even so, most simulation studies have simulated balanced scenarios, in which the sample size was equal across primary studies (e.g., Hedges et al., 2010; Van den Noortgate et al., 2013, 2015), while the reality is that the sample size can vary quite a lot across primary studies. Furthermore, in general the sample sizes selected in simulation studies are representative of the behavioral and social sciences and medical science field, where the median number of observations reported within studies is close to 100. However, the median sample size in studies that belong to biological sciences and cultural sciences is close to 30, which is smaller than the common value included in the simulation studies. Only the simulation studies of Hedges et al. (2010) and Tipton (2013, 2015) have considered such small values for the sample sizes.

Simulation studies have selected the values for the variance components in two different ways. Some studies have directly selected a specific value for the between-

studies and within-studies variances (e.g., Lee, 2014; Moeyaert et al., 2017), and other studies have used ratios to generate these factors (e.g., Park & Beretvas, 2018; Tipton, 2013). Among the first type of simulation studies, the total variance commonly used ranged from 0.10 to 0.30. The median total variance observed in this systematic review (that was obtained summing the medians of the variances at the second and third level) for Cohen's *d* ranged from 0.12 to 0.26 across disciplines, so the 'average' total variance is actually represented in the simulation studies. A similar range was observed for odds ratios (0.11–0.25). Regarding Pearson correlations, the total median variances were smaller and ranged from 0.034 to 0.19. Among the second type of simulation studies, we can see that, for instance, Park and Beretvas (2018) simulated two ratios for Cohen's *d:* 5:3:2 and 2:3:5, where the first number refers to the I^2 index of the sampling variance, the second number refers to the I^2 index of the second level variance, and finally the third number stands for the I^2 index of the third level variance. This last ratio is similar to the one found in this systematic review for behavioral and social sciences (i.e., was 2:2:6 and 2:3:5 for all types of effect sizes), but it is not representative of medical sciences (i.e., 1:4:5 for Cohen's *d*).

Finally, regarding the number of moderators, there are four simulation studies that test whether the moderator effect of only one variable (and its standard error) is correctly recovered. Only the Study 2 of Tipton (2015) explores the performance of the robust variance estimation method when up to four moderator variables are tested in the regression. However, our results indicate that at least 15.16% of the meta-analyses test several moderators at the same time. Specifically, from those meta-analyses that introduced at the same time all moderator variables, the average number of moderators tested was seven. Furthermore, none of the simulation studies explores the recovery of interaction effects, and 32 meta-analysis included in this systematic review do test for interaction effects. It is important that future simulation studies take into account these scenarios.

In sum, the application of multilevel models in the field of meta-analysis offers many opportunities, especially for the treatment of dependent effect sizes. However, researchers should start considering the application of multilevel models different from the three-level model if the meta-analytic data require it. Furthermore, we expect that researchers take into account the information reported in this systematic review to design future simulation studies in this field, so that their results can be generalized to real settings.

Author Notes This research has been supported by the Research Foundation – Flanders (FWO), through Grant G.0798.15N to the University of Leuven, Belgium. The opinion expressed are those of the authors and do not represent views of the FWO.

References

References marked with an asterisk (*) were included in systematic review

*Acar, S., Chen, X., & Cayirdag, N. (2018). Schizophrenia and creativity: A meta-analytic review. *Schizophrenia Research*, *195*, 23–31.

*Acar, S., & Sen, S. (2013). A multilevel meta-analysis of the relationship between creativity and schizotypy. *Psychology of Aesthetics, Creativity, and the Arts*, *7*, 214–228.

*Acar, S., Sen, S., & Cayirdag, N. (2016). Consistency of the performance and nonperformance methods in gifted identification: A multilevel meta-analytic review. *Gifted Child Quarterly*, *60*, 81–101.

*Appuhamy, J. A. D. R. N., Judy, J. V., Kebreab, E., & Kononoff, P. J. (2016). Prediction of drinking water intake by dairy cows. *Journal of Dairy Science*, *99*, 7191–7205.

*Assink, M., Van der Put, C. E., Hoeve, M., de Vries, S. L. A., Stams, G. J. J. M., & Oort, F. J. (2015). Risk factors for persistent delinquent behavior among juveniles: A meta-analytic review. *Clinical Psychology Review*, *42*, 47–61.

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*, 154–174.

*Baker, J. M., & Reiss, A. L. (2016). A meta-analysis of math performance in Turner syndrome. *Developmental Medicine and Child Neurology*, *58*, 123–130.

*Baron, A., Evangelou, M., Malmberg, L.-E., & Melendez-Torres, G.-J. (2015). *The Tools of the Mind curriculum for improving self-regulation in early childhood: A systematic review*. Oslow: The Campbell Collaboration.

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). Orl: Academic Press.

*Becker, D. J., Streicker, D. G., & Altizer, S. (2018). Using host species traits to understand the consequences of resource provisioning for host-parasite interactions. *Journal of Animal Ecology*, *87*, 511–525.

*Bentz, A. B., Becker, D. J., & Navara, K. J. (2016). Meta-analysis of yolk testosterone response to competition. *Royal Society Open Science*, *3*, 1–12.

*Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne social desirability scale. *Educational and Psychological Measurement*, *62*, 570–589.

*Bernstein, L. J., McCreath, G. A., Komeylian, Z., & Rich, J. B. (2017). Cognitive impairment in breast cancer survivors treated with chemotherapy depends on control group type and cognitive domains assessed: A multilevel meta-analysis. *Neuroscience and Biobehavioral Reviews*, *83*, 417–428.

*Blom, R., Kruyen, P. M., Van der Heijden, B. I. J. M., & Van Thiel, S. (2018). One HRM fits all? A meta-analysis of the effects of HRM practices in the public, semipublic, and private sector. *Review of Public Personnel Administration*, 1–33.

*Blum, D., & Holling, H. (2017). Spearman's law of diminishing returns. A meta-analysis. *Intelligence*, *65*, 60–66.

*Borg, J., Kiær, L. P., Lecarpentier, C., Goldringer, I., Gauffreteau, A., Saint-Jean, S., … Enjalbert, J. (2018). Unfolding the potential of wheat cultivar mixtures: A meta-analysis perspective and identification of knowledge gaps. *Field Crops Research*, *221*, 298–313.

*Bornmann, L, Mutz, R., & Daniel, H. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*, 226–238.

*Bornmann, Lutz, Mutz, R., & Daniel, H. -D. (2010). A Reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS One*, *5*, e14331.

*Bornmann, Lutz, Mutz, R., Hug, S. E., & Daniel, H.-D. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics*, *5*, 346–359.

*Bortolon, C., & Raffard, S. (2018). Self-face advantage over familiar and unfamiliar faces: A three-level meta-analytic approach. *Psychonomic Bulletin & Review*, *25*, 1287–1300.

*Bucher, O., Farrar, A. M., Totton, S. C., Wilkins, W., Waddell, L. A., Wilhelm, B. J., … Rajić, A. (2012). A systematic review-meta-analysis of chilling interventions and a meta-regression of various processing interventions for Salmonella contamination of chicken. *Preventive Veterinary Medicine*, *103*, 1–15.

*Bürkner, P.-C., Williams, D. R., Simmons, T. C., & Woolley, J. D. (2017). Intranasal oxytocin may improve high-level social cognition in schizophrenia, but not social cognition or neurocognition in general: A multilevel Bayesian meta-analysis. *Schizophrenia Bulletin*, *43*, 1291–1303.

*Carrasquilla-Henao, M., & Juanes, F. (2017). Mangroves enhance local fisheries catches: a global meta-analysis. *Fish and Fisheries*, *18*, 79–93.

*Carrillat, F. A., Legoux, R., & Hadida, A. L. (2018). Debates and assumptions about motion picture performance: a meta-analysis. *Journal of the Academy of Marketing Science*, *46*, 273–299.

*Chan, W. S., Levsen, M. P., & McCrae, C. S. (2018). A meta-analysis of associations between obesity and insomnia diagnosis and symptoms. *Sleep Medicine Reviews*, *40*, 170–182.

*Cheng, C., Cheung, M. W.-L., & Lo, B. C. Y. (2016). Relationship of health locus of control with specific health behaviors and global health appraisal: a meta-analysis and effects of moderators. *Health Psychology Review*, *10*, 460–477.

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

*Chita-Tegmark, M. (2016). Attention allocation in ASD: A review and meta-analysis of eye-tracking studies. *Review Journal of Autism and Developmental Disorders*, *3*, 209–223.

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

*Colagrossi, M., Rossignoli, D., & Maggioni, M. A. (2017). Does democracy cause growth? A meta-analysis perspective, 1–26.

*Compton, C. W. R., Heuer, C., Thomsen, P. T., Carpenter, T. E., Phyn, C. V. C., & McDougall, S. (2017). Invited review: A systematic literature review and meta-analysis of mortality and culling in dairy cattle. *Journal of Dairy Science*, *100*, 1–16.

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

*Curry, O. S., Rowland, L. A., Van Lissa, C. J., Zlotowitz, S., McAlaney, J., & Whitehouse, H. (2018). Happy to help? A systematic review and meta-analysis of the effects of performing acts of kindness on the well-being of the actor. *Journal of Experimental Social Psychology*, *76*, 320–329.

*De La Rue, L., Polanin, J. R., Espelage, D. L., & Pigott, T. D. (2014). *School-based interventions to reduce dating and sexual violence*. The Campbell Collaboration. Retrieved from https://campbellcollaboration.org/library/school-based-interventions-dating-and-sexual-violence.html

*De La Rue, L., Polanin, J. R., Espelage, D. L., & Pigott, T. D. (2017). A meta-analysis of school-based interventions aimed to prevent or reduce violence in teen dating relationships. *Review of Educational Research*, *87*, 7–34.

*De Noordhout, C. M., Devleesschauwer, B., Angulo, F. J., Verbeke, G., Haagsma, J., Kirk, M., Havelaar, A., Speybroeck, N. (2014). The global burden of listeriosis: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, *14*, 1073–1082.

*De Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input: Meta-analysis: Incidental L2 spoken word learning. *Language Learning*, *68*, 906–941.

*Vries, L. A. de. (2016). *The effectiveness of youth crime prevention* (Doctoral Dissertation). Retrieved from https://pure.uva.nl/ws/files/4506124/173543_thesis_ex_dankw_incl_cv.pdf

*Deb, S. K., Brown, D. R., Gough, L. A., Mclellan, C. P., Swinton, P. A., Andy Sparks, S., & Mcnaughton, L. R. (2018). Quantifying the effects of acute hypoxic exposure on exercise performance and capacity: A systematic review and meta-regression. *European Journal of Sport Science*, *18*, 243–256.

*Dekkers, T. J., Popma, A., Agelink van Rentergem, J. A., Bexkens, A., & Huizenga, H. M. (2016). Risky decision making in attention-deficit/hyperactivity disorder: A meta-regression analysis. *Clinical Psychology Review*, *45*, 1–16.

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

*Des Roches, S., Post, D. M., Turley, N. E., Bailey, J. K., Hendry, A. P., Kinnison, M. T., … Palkovacs, E. P. (2018). The ecological importance of intraspecific variation. *Nature Ecology and Evolution*, *2*, 57–64.

*Dijkstra, S., Creemers, H. E., Asscher, J. J., Deković, M., & Stams, G. J. J. M. (2016). The effectiveness of family group conferencing in youth care: A meta-analysis. *Child Abuse and Neglect*, *62*, 100–110.

*Dodell-Feder, D., & Tamir, D. I. (2018). Fiction reading has a small positive impact on social cognition: A meta-analysis. *Journal of Experimental Psychology: General*, *147*, 1713–1727.

*Dolan, E., Swinton, P. A., Sale, C., Healy, A., & O'Reilly, J. (2017). Influence of adipose tissue mass on bone mass in an overweight or obese population: Systematic review and meta-analysis. *Nutrition Reviews*, *75*, 858–870.

*Donnelly, S., Brooks, P. J., & Homer, B. D. (2015). *Examining the bilingual advantage on conflict resolution tasks: a meta-analysis* (p. 6). Presented at the Proceedings of the 37th Annual Meeting of the Cognitive Science Society, Austin, TX.

*Dopp, A. R., Borduin, C. M., White, M. H., & Kuppens, S. (2017). Family-based treatments for serious juvenile offenders: A multilevel meta-analysis. *Journal of Consulting and Clinical Psychology*, *85*, 335–354.

*Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., van den Hout, M. A., Baas, J. M. P. (2015). Review: Updated meta-analysis of fear conditioning in anxiety disorders. *Depression and Anxiety*, *32*, 239–253.

*Dworkin, E. R., Menon, S. V., Bystrynski, J., & Allen, N. E. (2017). Sexual assault victimization and psychopathology: A review and meta-analysis. *Clinical Psychology Review*, *56*, 65–81.

Fernández-Castilla, B., Maes, M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2018). A demonstration and evaluation of the use of cross-classified random-effects models for meta-analysis. *Behavior Research Methods,* 1–19.

*Fischer, R. (2013). Belonging, status, or self-protection? Examining justice motives in a three-level cultural meta-analysis of organizational justice effects. *Cross-Cultural Research*, *47*, 3–41.

*Fischer, R., & Boer, D. (2011). What is more important for national well-being: Money or autonomy? A meta-analysis of well-being, burnout, and anxiety across 63 societies. *Journal of Personality and Social Psychology*, *101*, 164–184.

*Fischer, R., & Derham, C. (2016). Is in-group bias culture-dependent? A meta-analysis across 18 societies. *SpringerPlus*, *5*, 1–9.

*Fischer, R., Hanke, K., & Sibley, C. G. (2012). Cultural and institutional determinants of social dominance orientation: a cross-cultural meta-analysis of 27 societies: SDO determinants at societal level. *Political Psychology*, *33*, 437–467.

*Fischer, R., & Mansell, A. (2009). Commitment across cultures: A meta-analytical approach. *Journal of International Business Studies*, *40*, 1339–1358.

*Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, *55*, 316–340.

*Fradkin, I., Strauss, A. Y., Pereg, M., & Huppert, J. D. (2018). Rigidly applied rules? revisiting inflexibility in obsessive compulsive disorder using multilevel meta-analysis. *Clinical Psychological Science*, *6*, 481–505.

*Francis, T. A. (2016). *The core components of cardiac rehabilitation for health-related quality of life in coronary heart disease patients: a systematic review and meta-analysis of randomized controlled trials* (Doctoral dissertation). University of Toronto.

*Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*, 296–321.

*Gao, S., Assink, M., Cipriani, A., & Lin, K. (2017). Associations between rejection sensitivity and mental health outcomes: A meta-analytic review. *Clinical Psychology Review*, *57*, 59–74.

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.

*Gilman, E., Chaloupka, M., Swimmer, Y., & Piovano, S. (2016). A cross-taxa assessment of pelagic longline by-catch mitigation measures: Conflicts and mutual benefits to elasmobranchs. *Fish and Fisheries*, *17*, 748–784.

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

*Gnambs, T. (2014). A meta-analysis of dependability coefficients (test–retest reliabilities) for measures of the Big Five. *Journal of Research in Personality*, *52*, 20–28.

*Gnambs, T. (2015). Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences*, *84*, 84–89.

*Gnambs, T., & Appel, M. (2018). Narcissism and social networking behavior: a meta-analysis: Narcissism and social networking sites. *Journal of Personality*, *86*, 200–212.

*Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: a meta-analysis. *Behavior Research Methods*, *47*, 1237–1259.

*Gnambs, T., & Kaspar, K. (2017). Socially desirable responding in web-based questionnaires: A meta-analytic review of the candor hypothesis. *Assessment*, *24*, 746–762.

*Goense, P. B., Assink, M., Stams, G.-J., Boendermaker, L., & Hoeve, M. (2016). Making 'what works' work: A meta-analytic study of the effect of treatment integrity on outcomes of evidence-based interventions for juveniles with antisocial behavior. *Aggression and Violent Behavior*, *31*, 106–115.

*Golivets, M., & Wallin, K. F. (2018). Neighbour tolerance, not suppression, provides competitive advantage to non-native plants. *Ecology Letters*, *21*, 745–759.

*Gubbels, J., van der Stouwe, T., Spruit, A., & Stams, G. J. J. M. (2016). Martial arts participation and externalizing behavior in juveniles: A meta-analytic review. *Aggression and Violent Behavior*, *28*, 73–81.

*Guy, N., Newton-Howes, G., Ford, H., Williman, J., & Foulds, J. (2018). The prevalence of comorbid alcohol use disorder in the presence of personality disorder: Systematic review and explanatory modelling: Alcohol use disorder prevalence in personality disorder: Systematic review. *Personality and Mental Health*, *12*, 216–228.

*Hanke, K., & Fischer, R. (2013). Socioeconomical and sociopolitical correlates of interpersonal forgiveness: A three-level meta-analysis of the Enright Forgiveness Inventory across 13 societies. *International Journal of Psychology*, *48*, 514–526.

*Hedger, N., Gray, K. L. H., Garner, M., & Adams, W. J. (2016). Are visual threats prioritized without awareness? A critical review and meta-analysis involving 3 behavioral paradigms and 2696 observers. *Psychological Bulletin*, *142*, 934–968.

Hedges, L. V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 37–47). New York: Russell Sage Foundation.

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

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.

*Hendriks, A. M., Van der Giessen, D., Stams, G. J. J. M., & Overbeek, G. (2018). The association between parent-reported and observed parenting: A multi-level meta-analysis. *Psychological Assessment*, *30*, 621–633.

*Hennessy, E. A., Tanner-Smith, E. E., & Steinka-Fry, K. T. (2015). Do brief alcohol interventions reduce tobacco use among adolescents and young adults? A systematic review and meta-analysis. *Journal of Behavioral Medicine*, *38*, 899–911.

*Hijbeek, R., van Ittersum, M. K., ten Berge, H. F. M., Gort, G., Spiegel, H., & Whitmore, A. P. (2017). Do organic inputs matter – a meta-analysis of additional yield effects for arable crops in Europe. *Plant and Soil*, *411*, 293–303.

*Hildebrand, M., Wibbelink, C. J. M., & Verschuere, B. (2018). Do impression management and self-deception distort self-report measures with content of dynamic risk factors in offender samples? A meta-analytic review. *International Journal of Law and Psychiatry*, *58*, 157–170.

*Holt, M. K., Vivolo-Kantor, A. M., Polanin, J. R., Holland, K. M., DeGue, S., Matjasko, J. L., Wolfe, M., Reid, G. (2015). Bullying and Suicidal Ideation and Behaviors: A Meta-Analysis. *Pediatrics*, *135*, e496–e509.

*Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*, 901–930.

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

Hox, J. J., & de Leeuw, E. D. (2003). Multilevel models for meta-analysis. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 90–111). Mahwah: Erlbaum.

*Jaffe, A. E., DiLillo, D., Hoffman, L., Haikalis, M., & Dykstra, R. E. (2015). Does it hurt to ask? A meta-analysis of participant reactions to trauma research. *Clinical Psychology Review*, *40*, 40–56.

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

*Karsay, K., Knoll, J., & Matthes, J. (2018). Sexualizing media use and self-objectification: A meta-analysis. *Psychology of Women Quarterly*, *42*, 9–28.

*Karwowski, M., & Lebuda, I. (2016). The big five, the huge two, and creative self-beliefs: A meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts*, *10*, 214–232.

*Kende, J., Phalet, K., Van den Noortgate, W., Kara, A., & Fischer, R. (2017). Equality revisited: A cultural meta-analysis of intergroup contact and prejudice. *Social Psychological and Personality Science*, 194855061772899.

*Kilgus, S. P., Eklund, K., Maggin, D. M., Taylor, C. N., & Allen, A. N. (2018). The student risk screening scale: A reliability and validity generalization meta-analysis. *Journal of Emotional and Behavioral Disorders*, *26*, 143–155.

*Klomp, J., & Valckx, K. (2014). Natural disasters and economic growth: A meta-analysis. *Global Environmental Change*, *26*, 183–195.

*Knapp, F., Viechtbauer, W., Leonhart, R., Nitschke, K., & Kaller, C. P. (2017). Planning performance in schizophrenia patients: a meta-analysis of the influence of task difficulty and clinical and sociodemographic variables. *Psychological Medicine*, *47*, 2002–2016.

*Knoll, J., & Matthes, J. (2017). The effectiveness of celebrity endorsements: a meta-analysis. *Journal of the Academy of Marketing Science*, *45*, 55–75.

*Knotter, M. H., Spruit, A., De Swart, J. J. W., Wissink, I. B., Moonen, X. M. H., & Stams, G. J. M. (2018). Training direct care staff working with persons with intellectual disabilities and challenging behaviour: A meta-analytic review study. *Aggression and Violent Behavior*, *40*, 60–72.

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

*Krieger, J. W. (2009). Single versus multiple sets of resistance exercise: a meta-regression: *Journal of Strength and Conditioning Research*, *23*, 1890–1901.

*Krieger, J. W. (2010). Single vs. multiple sets of resistance exercise for muscle hypertrophy: A meta-analysis: *Journal of Strength and Conditioning Research*, *24*, 1150–1159.

*Krieger, J. W., Sitren, H. S., Daniels, M. J., & Langkamp-Henken, B. (2006). Effects of variation in protein and carbohydrate intake on body mass and composition during energy restriction: a meta-regression. *The American Journal of Clinical Nutrition*, *83*, 260–274.

*Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: implications for theory and research. *British Educational Research Journal*, *36*, 807–830.

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

*Lee, C. I. S. G., Bosco, F. A., Steel, P., & Uggerslev, K. L. (2017). A metaBUS-enabled meta-analysis of career satisfaction. *Career Development International*, *22*, 565–582.

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

*Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., de Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin*, *144*, 394–425.

*Leijten, P., Melendez-Torres, G. J., Knerr, W., & Gardner, F. (2016). Transported versus homegrown parenting interventions for reducing disruptive child behavior: A multilevel meta-regression study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *55*, 610–617.

López-López, J. A., Page, M. J., Lipsey, M. W., & Higgins, J. P. T. (2018). Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Research Synthesis Methods, 9,* 336–351.

*Machts, N., Kaiser, J., Schmidt, F. T. C., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review*, *19*, 85–103.

*Maes, M. (2015). A reliability generalization study for a multidimensional loneliness scale. *European Journal of Psychological Assessment*, 10.

*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.

*Martineau, R., Ouellet, D. R., Kebreab, E., & Lapierre, H. (2016). Casein infusion rate influences feed intake differently depending on metabolizable protein balance in dairy cows: A multilevel meta-analysis. *Journal of Dairy Science*, *99*, 2748–2761.

*Martineau, R., Ouellet, D. R., Kebreab, E., White, R. R., & Lapierre, H. (2017). Relationships between postruminal casein infusion and milk production, and concentrations of plasma amino acids and blood urea in dairy cows: A multilevel mixed-effects meta-analysis. *Journal of Dairy Science*, *100*, 8053–8071.

*Matthes, J., Knoll, J., & von Sikorski, C. (2018). The "spiral of silence" revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression. *Communication Research*, *45*, 3–33.

*Mauger, C., Lancelot, C., Roy, A., Coutant, R., Cantisano, N., & Le Gall, D. (2018). Executive functions in children and adolescents with turner syndrome: A systematic review and meta-analysis. *Neuropsychology Review*, *28*, 188–215.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*, 114–140.

*Melendez-Torres, G. J., Dickson, K., Fletcher, A., Thomas, J., Hinds, K., Campbell, R., … Bonell, C. (2016). Positive youth development programmes to reduce substance use in young people: Systematic review. *International Journal of Drug Policy*, *36*, 95–103.

*Melendez-Torres, G. J., Tancred, T., Fletcher, A., Thomas, J., Campbell, R., & Bonell, C. (2018). Does integrated academic and health education prevent substance use? Systematic review and meta-analyses. *Child: Care, Health and Development*, *44*, 516–530.

Moeyaert, M., Ugille, M., Natasha Beretvas, S., 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.

*Montoya, R. M., Kershaw, C., & Prosser, J. L. (2018). A meta-analytic investigation of the relation between interpersonal attraction and enacted behavior. *Psychological Bulletin*, *144*, 673–709.

*Moore, F. R., Shuker, D. M., & Dougherty, L. (2016). Stress and sexual signaling: a systematic review and meta-analysis. *Behavioral Ecology*, *27*, 363–371.

*Musitelli, F., Romano, A., Møller, A. P., & Ambrosini, R. (2016). Effects of livestock farming on birds of rural areas in Europe. *Biodiversity and Conservation*, *25*, 615–631.

*Nettle, D., Andrews, C., & Bateson, M. (2017). Food insecurity as a driver of obesity in humans: The insurance hypothesis. *Behavioral and Brain Sciences*, *40*, 1–53.

*O'Mara, A. J., Marsh, H. W., & Craven, R. G. (2006). *A Comprehensive Multilevel Model Meta-Analysis of Self-Concept Interventions* (p. 12). Presented at the Fourth International Biennial SELF Research Conference, US.

*O'Shea, L. E., & Dickens, G. L. (2016). Performance of protective factors assessment in risk prediction for adults: Systematic review and meta-analysis. *Clinical Psychology: Science and Practice*, *23*, 126–138.

*Oosterhoff, M., Joore, M., & Ferreira, I. (2016). The effects of school-based lifestyle interventions on body mass index and blood pressure: A multivariate multilevel meta-analysis of randomized controlled trials. *Obesity Reviews*, *17*, 1131–1153.

*Op den Kelder, R., Van den Akker, A. L., Geurts, H. M., Lindauer, R. J. L., & Overbeek, G. (2018). Executive functions in trauma-exposed youth: a meta-analysis. *European Journal of Psychotraumatology*, *9*, 1450595.

*Orellano, P., Quaranta, N., Reynoso, J., Balbi, B., & Vasquez, J. (2017). Effect of outdoor air pollution on asthma exacerbations in children and adults: Systematic review and multilevel meta-analysis. *PLoS One*, *12*, e0174050.

*Ousey, G. C., & Kubrin, C. E. (2018). Immigration and crime: Assessing a contentious issue. *Annual Review of Criminology*, *1*, 63–84.

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

*Paauw, N. D., Terstappen, F., Ganzevoort, W., Joles, J. A., Gremmels, H., & Lely, A. T. (2017). Sildenafil during pregnancy. *Hypertension*, *70*, 998–1006.

*Paek, S. H., Abdulla, A. M., & Cramond, B. (2016). A meta-analysis of the relationship between three common psychopathologies-ADHD, anxiety, and depression-and indicators of little-c creativity. *Gifted Child Quarterly*, *60*, 117–133.

Park, S., & Beretvas, S. N. (2018). Synthesizing effects for multiple outcomes per study using robust variance estimation versus the three-level model. *Behavior Research Methods,* 1-20.

Pastor, D. A., & Lazowski, R. A. (2018). On the multilevel nature of meta-analysis: A tutorial, comparison of software programs, and discussion of analytic choices. *Multivariate Behavioral Research*, *53*, 74–89.

*Patinkin, Z. W., Feinn, R., & Santos, M. (2017). Metabolic consequences of obstructive sleep apnea in adolescents with obesity: A systematic literature review and meta-analysis. *Childhood Obesity*, *13*, 102–110.

*Pearce, L. J. (2017). *Children's true and false memories of valenced material* (Doctoral dissertation). Retrieved from: http://sro.sussex.ac.uk/id/eprint/68784

*Pfaller, J. B., Chaloupka, M., Bolten, A. B., & Bjorndal, K. A. (2018). Phylogeny, biogeography and methodology: a meta-analytic perspective on heterogeneity in adult marine turtle survival rates. *Scientific Reports*, *8*.

*Platt, L., Melendez-Torres, G. J., O'Donnell, A., Bradley, J., Newbury-Birch, D., Kaner, E., & Ashton, C. (2016). How effective are brief interventions in reducing alcohol consumption: do the setting, practitioner group and content matter? Findings from a systematic review and meta-regression analysis. *British Medical Journal Open*, *6*, e011473.

*Pratt, T. C., Turanovic, J. J., Fox, K. A., & Wright, K. A. (2014). Self-control and victimization: a meta-analysis: self-control and victimization. *Criminology*, *52*, 87–116.

*Pyrooz, D. C., Turanovic, J. J., Decker, S. H., & Wu, J. (2016). Taking stock of the relationship between gang membership and offending: A meta-analysis. *Criminal Justice and Behavior*, *43*, 365–397.

*Rabl, T., Jayasinghe, M., Gerhart, B., & Kühlmann, T. M. (2014). A meta-analysis of country differences in the high-performance work system–business performance relationship: The roles of national culture and managerial discretion. *Journal of Applied Psychology*, *99*, 1011–1041.

*Rapp, R. C., Van Den Noortgate, W., Broekaert, E., & Vanderplasschen, W. (2014). The efficacy of case management with persons who have substance abuse problems: A three-level meta-analysis of outcomes. *Journal of Consulting and Clinical Psychology*, *82*, 605–618.

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

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

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

*Reinhold, M., Bürkner, P.-C., & Holling, H. (2018). Effects of expressive writing on depressive symptoms-A meta-analysis. *Clinical Psychology: Science and Practice*, *25*, e12224.

*Roca Fraga, F. J., Lagisz, M., Nakagawa, S., Lopez-Villalobos, N., Blair, H. T., & Kenyon, P. R. (2018). Meta-analysis of lamb birth weight as influenced by pregnancy nutrition of multiparous ewes. *Journal of Animal Science*, *96*, 1962–1977.

*Romano, A., Costanzo, A., Rubolini, D., Saino, N., & Møller, A. P. (2017). Geographical and seasonal variation in the intensity of sexual selection in the barn swallow *Hirundo rustica* : a meta-analysis. *Biological Reviews*, *92*, 1582–1600.

*Romano, A., Saino, N., & Møller, A. P. (2017). Viability and expression of sexual ornaments in the barn swallow *Hirundo rustica* : A meta-analysis. *Journal of Evolutionary Biology*, *30*, 1929–1935.

*Roquet, R. F. (2017). *The influence of exercise on persistence of fear* (Doctoral dissertation). Retrieved from: https://repositories.lib.utexas.edu/handle/2152/63062

*Roquet, R. F., & Monfils, M.-H. (2018). Does exercise augment operant and Pavlovian extinction: A meta-analysis. *Journal of Psychiatric Research*, *96*, 73–93.

Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., & López-López, J. A. (2018). A methodological review of meta-analyses of the effectiveness of clinical psychology treatments. *Behavior Research Methods*, 1–17.

*Runge, M. S., Cheung, M. W.-L., & D'Angiulli, A. (2017). Meta-analytic comparison of trial- versus questionnaire-based vividness reportability across behavioral, cognitive and neural measurements of imagery. *Neuroscience of Consciousness*, 1–13.

*Saunders, B., Elliott-Sale, K., Artioli, G. G., Swinton, P. A., Dolan, E., Roschel, H., Sale, C., Gualano, B. (2017). β-alanine supplementation to improve exercise capacity and performance: a systematic review and meta-analysis. *British Journal of Sports Medicine*, *51*, 658–669.

*Schädel, C., Bader, M. K.-F., Schuur, E. A. G., Biasi, C., Bracho, R., Čapek, P., … Wickland, K. P. (2016). Potential carbon emissions dominated by carbon dioxide from thawed permafrost soils. *Nature Climate Change*, *6*, 950–953.

*Scharfen, J., Blum, D., & Holling, H. (2018). Response Time Reduction Due to Retesting in Mental Speed Tests: A Meta-Analysis. *Journal of Intelligence*, *6*, 1–28.

*Scharfen, J., Jansen, K., & Holling, H. (2018). Retest effects in working memory capacity tests: A meta-analysis. *Psychonomic Bulletin & Review*, *25*, 2175–2199.

*Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*, 44–66.

*Scheerens, J. (2007). *Review and meta-analyses of school and teaching effectiveness.* Twente, The Netherlands: University of Twente, Department of Educational Organization and Management.

*Schoenfeld, B., Aragon, A. A., & Krieger, J. W. (2013). The effect of protein timing on muscle strength and hypertrophy: A meta-analysis. *Journal of the International Society of Sports Nutrition*, *10:53*, 1–13.

*Schoenfeld, B. J., Aragon, A. A., & Krieger, J. W. (2015). Effects of meal frequency on weight loss and body composition: a meta-analysis. *Nutrition Reviews*, *73*, 69–82.

*Schoenfeld, B. J., Ogborn, D. I., & Krieger, J. W. (2015). Effect of Repetition Duration During Resistance Training on Muscle Hypertrophy: A Systematic Review and Meta-Analysis. *Sports Medicine*, *45*, 577–585.

*Schoenfeld, B. J., Wilson, J. M., Lowery, R. P., & Krieger, J. W. (2016). Muscular adaptations in low- versus high-load resistance training: A meta-analysis. *European Journal of Sport Science*, *16*, 1–10.

*Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., … Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin*, *144*, 343–393.

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

*Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin and Review*, *24*, 1077–1096.

*Spruit, A.(2017). *Keeping youth in play* (Doctoral dissertation). Retrieved from: https://pure.uva.nl/ws/files/8496595/07_References.pdf

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

*Spruit, A., Schalkwijk, F., van Vugt, E., & Stams, G. J. (2016). The relation between self-conscious emotions and delinquency: A meta-analysis. *Aggression and Violent Behavior*, *28*, 12–20.

*Spruit, A., van Vugt, E., van der Put, C., Van der Stouwe, T., & Stams, G.-J. (2016). Sports participation and juvenile delinquency: A meta-analytic review. *Journal of Youth and Adolescence*, *45*, 655–671.

*Steinka-Fry, K. T., Tanner-Smith, E. E., & Hennessy, E. A. (2015). Effects of Brief Alcohol Interventions on Drinking and Driving among Youth: A Systematic Review and Meta-analysis. *Journal of Addiction and Prevention*, *3,* 1–21.

*Steinmetz, H., Knappstein, M., Ajzen, I., Schmidt, P., & Kabst, R. (2016). How effective are behavior change interventions based on the theory of planned behavior? A three-level meta-analysis. *Zeitschrift Für Psychologie*, *224*, 216–233.

*Szumski, G., Smogorzewska, J., & Karwowski, M. (2017). Academic achievement of students without special educational needs in inclusive classrooms: A meta-analysis. *Educational Research Review*, *21*, 33–54.

*Tanner-Smith, E. E., & Risser, M. D. (2016). A meta-analysis of brief alcohol interventions for adolescents and young adults: variability in effects across alcohol measures. *The American Journal of Drug and Alcohol Abuse*, *42*, 140–151.

*Tanner-Smith, E. E., Steinka-Fry, K. T., Hennessy, E. A., Lipsey, M. W., & Winters, K. C. (2015). Can brief alcohol interventions for youth also address concurrent illicit drug use? Results from a meta-analysis. *Journal of Youth and Adolescence*, *44*, 1011–1023.

*Taylor, C. L. (2017). Creativity and mood disorder: A systematic review and meta-analysis. *Perspectives on Psychological Science*, *12*, 1040–1076.

*Teasdale, N., Elhussein, A., Butcher, F., Piernas, C., Cowburn, G., Hartmann-Boyce, J., … Scarborough, P. (2018). Systematic review and meta-analysis of remotely delivered interventions using self-monitoring or tailored feedback to change dietary behavior. *The American Journal of Clinical Nutrition*, *107*, 247–256.

*Ter Beek, E., Kuiper, C. H. Z., van der Rijken, R. E. A., Spruit, A., Stams, G. J. J. M., & Hendriks, J. (2018). Treatment effect on psychosocial functioning of juveniles with harmful sexual behavior: A multilevel meta-analysis. *Aggression and Violent Behavior*, *39*, 116–128.

*Ter Beek, E., Spruit, A., Kuiper, C. H. Z., van der Rijken, R. E. A., Hendriks, J., & Stams, G. J. J. M. (2018). Treatment effect on recidivism for juveniles who have sexually offended: A Multilevel meta-analysis. *Journal of Abnormal Child Psychology*, *46*, 543–556.

*Thompson, T., Oram, C., Correll, C. U., Tsermentseli, S., & Stubbs, B. (2017). Analgesic effects of alcohol: A systematic review and meta-analysis of controlled experimental studies in healthy participants. *The Journal of Pain*, *18*, 499–510.

*Tingir, S., Cavlazoglu, B., Caliskan, O., Koklu, O., & Intepe-Tingir, S. (2017). Effects of mobile devices on K-12 students' achievement: a meta-analysis: Effects of mobile devices. *Journal of Computer Assisted Learning*, *33*, 355–369.

Tipton, E. (2013). Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods*, *4*, 169–187.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*, 375–393.

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

*Van Aar, J., Leijten, P., Orobio de Castro, B., & Overbeek, G. (2017). Sustained, fade-out or sleeper effects? A systematic review and meta-analysis of parenting interventions for disruptive child behavior. *Clinical Psychology Review*, *51*, 153–163.

*Van Cleemput, E., Vanierschot, L., Fernández-Castilla, B., Honnay, O., & Somers, B. (2018). The functional characterization of grass- and shrubland ecosystems using hyperspectral remote sensing: trends, accuracy and moderating variables. *Remote Sensing of Environment*, *209*, 747–763.

*Van Dam, L., Smit, D., Wildschut, B., Branje, S. J. T., Rhodes, J. E., Assink, M., & Stams, G. J. J. M. (2018). Does natural mentoring matter? A multilevel meta-analysis on the association between natural mentoring and youth outcomes. *American Journal of Community Psychology*, *62*, 203–220.

*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., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*, 576–594.

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

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

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

*Van der Hallen, R. (2016). *Little things, big things: Perceptual organization in children with and without ASD* (Doctoral dissertation). Retrieved from Lirias, Source ID LIRIAS1769040.

*Van der Hallen, R., Evers, K., Brewaeys, K., Van den Noortgate, W., & Wagemans, J. (2015). Global processing takes time: A meta-analysis on local–global visual processing in ASD. *Psychological Bulletin*, *141*, 549–573.

*Van der Pol, T. M., Hoeve, M., Noom, M. J., Stams, G. J. J. M., Doreleijers, T. A. H., van Domburgh, L., & Vermeiren, R. R. J. M. (2017). Research Review: The effectiveness of multidimensional family therapy in treating adolescents with multiple behavior problems - a meta-analysis. *Journal of Child Psychology and Psychiatry*, *58*, 532–545.

*Van der Put, C. E., Assink, M., & Boekhout van Solinge, N. F. (2017). Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments. *Child Abuse and Neglect*, *73*, 71–88.

*Van der Put, C. E., Assink, M., Gubbels, J., & Boekhout van Solinge, N. F. (2018). Identifying effective components of child maltreatment interventions: A Meta-analysis. *Clinical Child and Family Psychology Review*, *21*, 171–202.

*Van Kesteren, C. F. M. G., Gremmels, H., de Witte, L. D., Hol, E. M., Van Gool, A. R., Falkai, P. G., … Sommer, I. E. C. (2017). Immune involvement in the pathogenesis of schizophrenia: a meta-analysis on postmortem brain studies. *Translational Psychiatry*, *7*, e1075.

Van Lissa, C. J. (2017). MetaForest: Exploring heterogeneity in meta-analysis using random forests. https://doi.org/10.31234/osf.io/myg6s

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

*Vonderlin, R., Kleindienst, N., Alpers, G. W., Bohus, M., Lyssenko, L., & Schmahl, C. (2018). Dissociation in victims of childhood abuse or neglect: A meta-analytic review. *Psychological Medicine*, *48*, 2467–2476.

*Voyer, D., Voyer, S. D., & Tramonte, L. (2012). Free-viewing laterality tasks: A multilevel meta-analysis. *Neuropsychology*, *26*, 551–567.

*Vukasović, T., & Bratko, D. (2015). Heritability of personality: A meta-analysis of behavior genetic studies. *Psychological Bulletin*, *141*, 769–785.

*Weiss, M., Hoegl, M., & Gibbert, M. (2017). How does material resource adequacy affect innovation project performance? A meta-

analysis: material resource adequacy and innovation project performance. *Journal of Product Innovation Management*, *34*, 842–863.

*Weisz, J. R., Kuppens, S., Eckshtain, D., Ugueto, A. M., Hawley, K. M., & Jensen-Doss, A. (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care: A multilevel meta-analysis. *Journal of the American Medical Association of Psychiatry*, *70*, 750.

*Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., … Fordwood, S. R. (2017). What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. *American Psychologist*, *72*, 79–117.

*Welmers-van de Poll, M. J., Roest, J. J., van der Stouwe, T., van den Akker, A. L., Stams, G. J. J. M., Escudero, V., … de Swart, J. J. W. (2018). Alliance and treatment outcome in family-involved treatment for youth problems: A three-level meta-analysis. *Clinical Child and Family Psychology Review*, *21*, 146–170.

*White, R. L., Babic, M. J., Parker, P. D., Lubans, D. R., Astell-Burt, T., & Lonsdale, C. (2017). domain-specific physical activity and mental health: A meta-analysis. *American Journal of Preventive Medicine*, *52*, 653–666.

*Wibbelink, C. J. M., Hoeve, M., Stams, G. J. J. M., & Oort, F. J. (2017). A meta-analysis of the association between mental disorders and juvenile recidivism. *Aggression and Violent Behavior*, *33*, 78–90.

*Wilkins, W., Rajić, A., Parker, S., Waddell, L., Sanchez, J., Sargeant, J., & Waldner, C. (2010). Examining heterogeneity in the diagnostic accuracy of culture and pcr for salmonella spp. in swine: a systematic review/meta-regression approach: Systematic review: Salmonella tests in pigs. *Zoonoses and Public Health*, *57*, 121–134.

*Williamson, V., Creswell, C., Fearon, P., Hiller, R. M., Walker, J., & Halligan, S. L. (2017). The role of parenting behaviors in childhood post-traumatic stress disorder: A meta-analytic review. *Clinical Psychology Review*, *53*, 1–13.

*Yeager, D. S., Fong, C. J., Lee, H. Y., & Espelage, D. L. (2015). Declines in efficacy of anti-bullying programs among older adolescents: Theory and a three-level meta-analysis. *Journal of Applied Developmental Psychology*, *37*, 36–51.

*Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education*, *31*, 412–422.

*Yu, J., Lim, H.-Y., Abdullah, F. N. D. M., Chan, H.-M., Mahendran, R., Ho, R., … Feng, L. (2018). Directional associations between memory impairment and depressive symptoms: data from a longitudinal sample and meta-analysis. *Psychological Medicine*, *48*, 1664–1672.

*Zeegers, M. A. J., Colonnesi, C., Stams, G.-J. J. M., & Meins, E. (2017). Mind matters: A meta-analysis on parental mentalization and sensitivity as predictors of infant–parent attachment. *Psychological Bulletin*, *143*, 1245–1272.

*Zetsche, U., Bürkner, P.-C., & Schulze, L. (2018). Shedding light on the association between repetitive negative thinking and deficits in cognitive control – A meta-analysis. *Clinical Psychology Review*, *63*, 56–65.

Open Practices Statements The data of the characteristics of meta-analysis with more than one random effect are available at https://osf.io/znc68/. This study was not preregistered.