2016 ■ Vol. 12 ■ no. 3

Fitting three-level meta-analytic models in R: A step-by-step tutorial

Mark Assink[a, ✉] and Carlijn J. M. Wibbelink[b]

[a]Research Institute of Child Development and Education; University of Amsterdam
[b]Psychology Research Institute; University of Amsterdam

Abstract ■ Applying a multilevel approach to meta-analysis is a strong method for dealing with dependency of effect sizes. However, this method is relatively unknown among researchers and, to date, has not been widely used in meta-analytic research. Therefore, the purpose of this tutorial was to show how a three-level random effects model can be applied to meta-analytic models in R using the rma.mv function of the metafor package. This application is illustrated by taking the reader through a step-by-step guide to the multilevel analyses comprising the steps of (1) organizing a data file; (2) setting up the R environment; (3) calculating an overall effect; (4) examining heterogeneity of within-study variance and between-study variance; (5) performing categorical and continuous moderator analyses; and (6) examining a multiple moderator model. By example, the authors demonstrate how the multilevel approach can be applied to meta-analytically examining the association between mental health disorders of juveniles and juvenile offender recidivism. In our opinion, the rma.mv function of the metafor package provides an easy and flexible way of applying a multi-level structure to meta-analytic models in R. Further, the multilevel meta-analytic models can be easily extended so that the potential moderating influence of variables can be examined.

Keywords ■ meta-analysis, multilevel analysis. **Tools** ■ R, rma.mv, metafor.

✉ M.Assink@UvA.nl

MA: na; *CJMW*: na

10.20982/tqmp.12.3.p154

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer.

Introduction

The term meta-analysis refers to a stepwise procedure and a set of statistical techniques for combining results of independent primary studies, so that overall conclusions regarding a specific topic can be drawn. In general, the meta-analytic process can be divided into the following steps: (1) formulating a research problem; (2) searching for relevant primary studies; (3) retrieving information from the primary studies; (4) integrating the retrieved information in statistical analyses; and (5) interpreting the results from the analyses and drawing overall conclusions. In this tutorial, we specifically focus on the statistical analyses in meta-analytic research (the fourth step mentioned above). Throughout the years, a large number of books have been written on meta-analysis and for a comprehensive overview of all aspects involved in meta-analytic research, we refer the reader to the work of Borenstein, Hedges, Higgins, and Rothstein (2009), Cooper (2010), Hunter and Schmidt (2004), Lipsey and Wilson (2001) and Mullen (1989).

After a research problem has been formulated and the search procedure for relevant primary studies has been finished, it is time for the research synthesist to retrieve information from all primary studies in a coding procedure. In essence, there are two aspects to coding studies: coding information about empirical findings reported in primary studies that can be expressed in effect sizes (i.e., the dependent variable), and the coding of factors, such as study design, ethnicity of the sample, and type of instruments used, that may influence the nature and magnitude of the empirical findings (i.e., the independent variables) (Lipsey & Wilson, 2001). For integrating empirical findings reported in primary studies, it is necessary that each empirical finding

on a topic of interest is expressed in an effect size, which Cohen (1988) has defined as a quantitative indication of the *degree to which [a] phenomenon is present in the population* (pp. 9 – 10). The larger the value, the greater the degree to which a phenomenon is present, or in other words, the larger the effect. Common metrics for effect size are the standardized difference between the mean of two different groups (Cohen's d), the correlation coefficient (r or Fisher's Z when transformed), and the odds-ratio.

An important requirement in traditional univariate meta-analytic approaches is that there is no dependency between effect sizes in the data set that is to be analyzed (e.g., Rosenthal, 1984). If there is dependency between effect sizes (i.e., effect sizes are correlated), there is overlap in information to which correlated effect sizes are referring to. In this way the available information is 'inflated' and consequently leads to an overconfidence in the results of a meta-analysis (Van den Noortgate, Lòpez-Lòpez, Marìn-Martìnez, & Sànchez-Meca, 2013). Lipsey and Wilson (2001) emphasize that for meeting the requirement of non-independency, only one effect size per primary study should be included. After all, it is likely that effect sizes extracted from the same study are more alike (and thus interdependent) than effect sizes extracted from different studies, because the former may be based on the same participants, instruments, and/or circumstances in which the research was conducted (Houben, Van den Noortgate, & Kuppens, 2015).

Different solutions for dealing with dependency of effect sizes have been described in the literature (see, for instance, Borenstein et al., 2009; Cooper, 2010; Del Re, 2015; Hedges & Olkin, 1985; Lipsey & Wilson, 2001; Rosenthal, 1984; Schmidt & Hunter, 2015). Common methods for handling dependency of effect sizes are: simply ignoring the dependency and analyzing the effect sizes as if they were independent; averaging the dependent effect sizes within studies into a single effect size by calculating an unweighted or - less biased - weighted average; selecting only one effect size per study (also referred to as eliminating effect sizes); and shifting the unit of analysis meaning that one unit of analysis is selected after which effect sizes are averaged within each unit. Some of these methods are quite conservative, whereas others produce more accurate effect sizes. Cheung (2015) presents a more detailed overview of these strategies and their limitations in his book on applying a structural equation modeling approach to meta-analysis.

When averaging or eliminating effect sizes in primary studies, there may not only be the problem of a lower statistical power in the analyses due to information loss, but also the problem of a limit in the research questions that can be addressed in a meta-analytic research project (Cheung, 2015). After all, informative differences between effect sizes are lost and can no longer be identified in the analyses. In addition, Cheung notes that extracting a single effect size from each primary study implies that homogeneity of effect sizes within studies is assumed, which is, in most instances, a questionable assumption. By stepping away from the traditional univariate approach to meta-analysis, it becomes possible to deal with dependency of effect sizes in such a way that a research synthesist can extract all relevant effect sizes from each primary study without needing to reduce the number of effect sizes in any way. By performing the analyses using all relevant effect sizes, all information can be preserved and maximum statistical power can be achieved. In addition, there is no assumption of homogeneity of effect sizes within studies.

Applying a three-level structure to a meta-analytic model (Cheung, 2014; Hox, 2010; Van den Noortgate et al., 2013, 2014) is a better approach for dealing with dependency of effect sizes than the methods just mentioned. This three-level meta-analytic model considers three different variance components distributed over the three levels of the model: sampling variance of the extracted effect sizes at level 1; variance between effect sizes extracted from the same study at level 2; and variance between studies at level 3. In short, this model allows effect sizes to vary between participants (level 1), outcomes (level 2), and studies (level 3). Contrary to several other statistical techniques, the multilevel approach does not require the correlations between outcomes reported within primary studies to be known for estimating the covariance matrix of the effect sizes, since the second level in the above described three-level meta-analytic model accounts for sampling covariation (Van den Noortgate et al., 2013). Because (estimates of) correlations between outcomes are rarely reported in primary studies and therefore difficult to obtain, the use of multilevel models in meta-analytic research is a very practical way to account for interdependency of effect sizes. Further, the three-level approach allows examining differences in outcomes within studies (i.e., within-study heterogeneity) as well as differences between studies (i.e., between-study heterogeneity). If there is evidence for heterogeneity in effect sizes, moderator analyses can be conducted to test variables that may explain within-study or between-study heterogeneity. For these analyses, the three-level random effects model can easily be extended with study and effect size characteristics, making the model a three-level mixed effects model.

Despite the fact that using multilevel modeling in meta-analysis is a strong method for dealing with interdependency of effect sizes, it is a rather unknown method among scholars and has not been widely applied yet in meta-analytic research. Therefore, the main purpose of this tu-

torial is to show how the above described three-level structure can be applied to meta-analytic models. For this purpose, we use the rma.mv function of the metafor package (Viechtbauer, 2015), which can be invoked in the statistical software environment R (R Development Core Team, 2016). The metafor package was written by Wolfgang Viechtbauer and comprises a large set of functions for conducting meta-analyses. One of the many features of this flexible R package is that it allows users to fit a variety of meta-analytic models in which different approaches to analysis can be used. The rma.mv function is part of this package and makes it possible to fit multilevel meta-analytic models that can be extended by including moderators. To illustrate how a three-level random effects meta-analytic model can be set up using the rma.mv function in the R environment, we will present an example of meta-analytic research on the association between mental health disorders and juvenile offender recidivism, which was adapted from the work of Wibbelink, Hoeve, Stams, and Oort (2016). The reader will be guided through this example in a stepwise manner. First, we will illustrate how a data set should be organized and how the R environment should be set up. Second, we will demonstrate how an overall effect can be estimated using a three-level meta-analytic model. Third, we will discuss within-study heterogeneity as well as between-study heterogeneity, and fourth, we will illustrate the steps that are involved in performing a moderator analysis. Lastly, we will show how moderators can be analyzed jointly in one multiple moderator model, in order to examine the unique contribution of moderators.

Example: The association between mental health disorders of juveniles and juvenile offender recidivism

In their meta-analytic study, Wibbelink et al. (2016) focused on associations between mental health disorders of delinquent juveniles and subsequent delinquent behavior of those juveniles (i.e., recidivism). More specifically, the first aim of the study was to meta-analytically estimate an overall association between mental health disorders of juveniles and recidivism, since there are considerable differences in the associations found in primary studies. By statistically summarizing primary studies, better insight is provided in the true association between mental health disorders of juveniles and recidivism. Because primary studies differ from each other in several ways (e.g., differences in the way recidivism is defined, differences in assessing recidivism, and differences in methodological characteristics), a second aim of the study was to examine whether (and how) the association between mental health disorders of juveniles and recidivism is moderated by a number of variables. For the present tutorial, we used a subset of the data set that Wibbelink and colleagues used in their meta-analytic study.

Organizing the data file

Prior to analyzing the effect sizes in a data set, it is first important to properly organize a data file, so that the three-level meta-analytic models can be built in the R environment. An excerpt of the data file that is used in the example described in the present tutorial is shown in Table 1. From this table, it can be derived that each row represents one effect size extracted from one primary study. The first four columns from the left represent the variables that are mandatory to create in order to properly build the three-level meta-analytic models. In the first column, each independent study is designated with a unique identifier in the variable `studyID`, and in the second column, each extracted effect size is designated with a unique identifier in the variable `effectsizeID`. As can be seen in the table, six effect sizes were extracted from study 1, three effect sizes from study 2, six effect sizes from study 3, one effect size from study 11, one effect size from study 12, and two effect sizes from study 16. The variable labeled `y` contains all actual effect sizes, and in this example, all effects are expressed in Cohen's d (but other metrics for the effect size, such as Fisher's z, can also be analyzed with the rma.mv function of the metafor package). Each effect size represents the difference in recidivism rates between juveniles with a mental health disorder and a comparison group of juveniles without a mental health disorder. A positive value of Cohen's d indicates that the prevalence of recidivism is higher in the group of juveniles with a mental health disorder relative to the comparison group, whereas a negative value of Cohen's d is indicative of the opposite. According to the criteria formulated by Cohen (1988), d values of .2, .5, and .8 can be interpreted as small, moderate, and large effects, respectively. The variable labeled `v` contains the sampling variance that corresponds with the observed effect size in the variable `y` and can be obtained by squaring the standard error.

The other variables that are part of the data set are tested in moderator analyses as potential moderators of the overall association between juveniles with a mental health disorder and recidivism. In our example, the potential moderators that will be examined are (1) publication status of the primary study; (2) type of delinquent behavior in which juveniles have recidivated; and (3) the year in which a primary study was published. Prior to testing categorical variables as potential moderators of the overall effect, we created a dummy variable for each category of a categorical variable (see Table 1). At first glance, it may seem redundant to create a dummy variable for each of the categories rather than for only the categories that are tested against a reference category (i.e., total number

Table 1 ■ Excerpt of the Data Set Used in the Present Example.

studyID	effectsizeID	y	v	pstatpub	pstatnotpub	typegen	typeovert	typecovert	pyear
1	1	.9066	.0740	1	0	1	0	0	5
1	2	.4295	.0398	1	0	1	0	0	5
1	3	.2679	.0481	1	0	1	0	0	5
1	4	.2078	.0239	1	0	1	0	0	5
1	5	.0526	.0331	1	0	1	0	0	5
1	6	-.0507	.0886	1	0	1	0	0	5
2	7	.5117	.0115	1	0	1	0	0	2
2	8	.4738	.0076	1	0	1	0	0	2
2	9	.3544	.0065	1	0	1	0	0	2
3	10	2.2844	.3325	1	0	1	0	0	-9
3	11	2.1771	.3073	1	0	1	0	0	-9
3	12	1.7777	.2697	1	0	1	0	0	-9
3	13	1.5480	.4533	1	0	1	0	0	-9
3	14	1.4855	.1167	1	0	1	0	0	-7
3	15	1.4836	.1706	1	0	1	0	0	-7
11	59	.8615	.1591	0	1	0	1	0	-7
12	63	.2994	.0041	0	1	0	1	0	5
16	93	-.5675	.0340	1	0	0	0	1	3
16	94	-.7586	.0437	1	0	0	0	1	3

Note. The data set used in the present example was based on the data set created by Wibbelink, Hoeve, Stams, and Oort (2016). studyID = Unique identifier for each primary study; effectsizeID = Unique identifier for each effect size; y = Variable containing all effect sizes; v = Variable containing all sampling variances; pstatpub = Published primary studies (0 = not published; 1 = published); Pstatnotpub = Unpublished primary studies (0 = published; 1 = unpublished); typegen = General delinquency; typeovert = Overt delinquency; typecovert = Covert delinquency. All variables are explained in the text.

of categories – 1). However, we were not only interested in the mean effect of a reference category, but also in the mean effect (including significance and confidence interval) of the other categories that are tested against a reference category. In order to obtain these results, we created a dummy variable for each category of a discrete variable that is tested as a potential moderator. We will further elaborate on this in the section on moderator analyses. So, in our example, we created two dummy variables for publication status and three dummy variables for type of delinquency. In the dichotomous variable `pstatpub`, it was coded whether a primary study was published or not (1 = published; 0 = not published). The dichotomous variable `pstatnotpub` was created by inverting the values of the variable `pstatpub`, so that 0 is indicative of a published study and 1 is indicative of an unpublished study. Both variables are mutually exclusive, as can be seen in Table 1. In the variables `typegen` (i.e., general delinquent behavior), `typeovert` (i.e., overt delinquent behavior), and `typecovert` (i.e., covert delinquent behavior). The value 1 in these three dummy variables is indicative of the specific type of delinquency being applicable, whereas the value 0 indicates that the specific type of delinquency is not applicable. Once again, these dummy variables are mutually exclusive. The publication year of a study was regarded as a continuous variable and after the publication year of all primary studies was coded, the variable was centered around its mean. The results were stored in the variable `pyear` (see Table 1). Prior to the analyses (but not visible in Table 1), it was checked whether outlying effect sizes were present in the data set by screening for standardized z values larger than 3.29 or smaller than -3.29 (Tabachnik & Fidell, 2013). In case of missing values in the variables that were to be tested as potential moderators, the cells were left empty (i.e., system missing values which are not visible in Table 1). Note that the data set used in the present example can be downloaded as a comma separated values file (named `dataset.csv`) from the journal's website.

Setting up the R environment

The statistical software environment R (we recommend at least version 3.2.2) can be downloaded from the following websites:
http://cran.r-project.org/bin/windows/base/ (for Windows);
http://cran.r-project.org/bin/macosx/ (for OS X).

R provides a basic graphical user interface, but it is rather easy to install a more productive developmental environment for R (such as RStudio), if desired by the user. After installing R, the user needs to define a working directory in which syntax, data, and other files can be found by the R environment. This can be done by running the syntax in Listing 1. Note that all syntax should be entered at the command prompt (>) of the R environment and that all text after a number sign (#) is considered a comment and will not be executed by R. Readers who are interested in replicating our analyses can therefore leave out the comments in the syntaxes presented in this tutorial.

Next, the user needs to install and load the metafor package that comprises the rma.mv function, which will be invoked later on for building the multilevel meta-analytic model. Installing and loading the metafor package can be performed by running the syntax in Listing 2.

Next, the data set needs to be imported into the R environment. Since our data was saved in the file `dataset.csv`, which is in the comma delimited format, we need to import this file by running the syntax in Listing 3.

In order to check whether the data was correctly imported in the R environment, the user can screen the imported data by invoking several functions in a sequential order (see the syntax in Listing 4).

Calculating an overall effect

First, the overall association between juveniles with a mental health disorder and recidivism (i.e., the overall effect) will be estimated by fitting a three-level meta-analytic model to the data that will only consist of an intercept representing the overall effect. For this purpose, we use the rma.mv function of the metafor package, by running the syntax in Listing 5.

Below, we will first take a closer look on the elements of the syntax in Listing 5 that are taken as arguments by the `rma.mv` function.

- `overall` = the name of the object in which the results of the `rma.mv` function will be stored. In our example, we have named this object `overall`, since we are first estimating an overall effect;
- `y` = the name of the variable containing all effect sizes (which are Cohen's d values in the present example);
- `v` = the name of the variable containing all sampling variances;
- `random` = the argument that is taken by the `rma.mv` function when the user wants to perform a random-effects meta-analysis. Because the primary studies in the present meta-analytic example were considered to be a random sample of the population of studies, we wanted to perform a random-effects meta-analysis by invoking the `rma.mv` function with the `random` argument (for more information on the random-effects approach, see for instance Raudenbush (2009), Van den Noortgate and Onghena (2003).
- `list(~ 1 | effectsizeID, ~ 1 | studyID)` = the element needed for defining the three-level structure of the meta-analytic model. `effectsizeID` (i.e., the variable containing the unique identifiers of all effect sizes in the data set) defines the second level of the three-level model at which the variance between effect sizes within primary studies is distributed. `studyID` (i.e., the variable containing the unique identifiers of all primary studies in the data set) defines the third level of the three-level model at which the variance between studies is distributed. For both grouping variables (i.e., `effectsizeID` and `studyID`) accounts that the same random effect is assigned to effect sizes with the same value of the grouping variable (i.e., effect sizes are not assumed to be independent), whereas different random effects are assigned to effect sizes having different values of the grouping variable (i.e., effect sizes are assumed to be independent). In this syntax element, the random effects variance is denoted by `~ 1` and is assigned to a grouping variable by the vertical bar (i.e., `|`). Note that the first level of the model at which the sampling variance of all extracted effect sizes is distributed, does not need to be defined in the syntax. The sampling variance is not estimated in the meta-analytic model and is considered to be known. In this example, we will use the formula as given by Cheung (2014, pg. 2015) to estimate the sampling variance parameter at the first level of the model, and we will return on this issue later on.
- `tdist=TRUE` = the argument specifying that test statistics and confidence intervals must be based on the t-distribution. See below for more information on this argument.
- `data=dataset` = the argument describing which object contains the data set.

We will now take a closer look at the `tdist=TRUE` argument of the syntax. The default settings of the `rma.mv` function prescribe that test statistics of individual coefficients and confidence intervals are based on the normal distribution (i.e., the Z distribution). Further, the omnibus test used for testing multiple coefficients in a meta-analytic model that is extended with potential moderating variables is, by default, based on the chi-square distribution with m degrees of freedom (m = number of coefficients tested in the model, excluding the intercept, if present in the model). Several scholars showed that using the Z distribution in assessing the significance of model coefficients and in building confidence intervals around these coeffi-

Listing 1 ■ Setting the Working Directory.

```
# Setting the working directory;
# Mind the forward slashes in the syntax.
setwd("C:/research/meta-analysis/data")
```

Listing 2 ■ Installing and Loading the Metafor Package.

```
# Installing and loading the metafor package.
install.packages("metafor")
library(metafor)
```

cients, may lead to an increase in the number of unjustified significant results (see, for instance, Li, Shi, & Roth, 1994; Ziegler, Koch, & Victor, 2001). To reduce this problem, the user can apply the Knapp and Hartung (2003) adjustment to the analyses by passing the argument `tdist=TRUE` to the `rma.mv` function.

By applying the Knapp and Hartung's (2003) adjustment, the calculation of standard errors, p values, and confidence intervals is slightly modified. To be precise, test statistics of individual coefficients will be based on the t distribution with k (number of effect sizes) – p (total number of coefficients in the model including the intercept) degrees of freedom. If an omnibus test is performed (only relevant when testing potential moderating variables by extending the intercept-only model with predictors), it will be based on the F distribution in which the degrees of freedom of the numerator (df1) equals the number of coefficients in the model, and in which the degrees of freedom of the denominator (df2) equals k (number of effect sizes) – p (total number of coefficients in the model including the intercept). In case the intercept-only model is extended with only one predictor, the F value of the omnibus test equals the square of the t value associated with the regression coefficient of the predictor. The studies of Assink et al. (2015), Houben et al. (2015) and Weisz et al. (2013) are examples of published meta-analytic research in which the Knapp and Hartung adjustment is applied. As for calculating the degrees of freedom, a Satterthwaite correction (Satterthwaite, 1946) is sometimes applied when there are differences in variances of the groups that are to be compared. This often results in fractional degrees of freedom (see, for instance, Table 2 in the work of Weisz et al., 2013 and Table 2 in the work of Houben et al., 2015). This Satterthwaite correction is not (yet) available in the `rma.mv` function, and therefore it cannot be applied when there are differences in variances between groups. However, until now, this does not seem problematic, since there is no empirical evidence available showing that applying the Satterthwaite correction produces more robust results in meta-analytic models (Viechtbauer, 2015, personal communication).

The results of fitting a three-level intercept only model to the data can be printed on screen by running the syntax in Listing 6. Running this syntax will produce the output that is shown in Output 1.

We will now proceed with a detailed explanation of Output 1.

- `k = 100; method: REML` implies that the data set comprises 100 effect sizes (i.e., 100 rows in the data set) and that the REstricted Maximum Likelihood estimation method (REML) is used for estimating the parameters in the model. It is often possible to choose between different estimation methods in statistical software, and each estimation method has its own advantages and disadvantages. The REML method is in some ways superior to other methods (see, for instance, Hox, 2010; Viechtbauer, 2005), but has also restrictions (e.g., Cheung, 2014; Van den Noortgate et al., 2013). In this tutorial, we will not further discuss this issue. However, it is important to note that by using the REML method, it is not possible to perform a log-likelihood-ratio test to compare the fit of an intercept-only model (i.e., a model without predictors) to a model with predictors (Hox, 2010; Van den Noortgate et al., 2014, for more information, see).
- `Loglik`, `Deviance`, `AIC`, `BIC`, `AICc` are goodness-of-fit indices for the meta-analytic model and provide information on how well the model fits the data set. In this tutorial, we will not further discuss the technical details of these indices.
- As for the variance components, it can be derived from the output that `0.112` is the estimated value for the variance between effect sizes within studies (distributed at the second level of the model) and that `0.188` is the estimated value for the variance between studies (distributed at the third level of the model). The results in the columns `nlvls` and `factor` tell us that the data set comprises 100 effect sizes (factor

Listing 3 ■ Importing the Data in R.

```
# Importing data saved in a comma separated values (CSV) file;
# The data file to be imported was named "dataset.csv";
# All data saved in the file "dataset.csv" is read by invoking
# the read.csv function and assigned to a newly created object
# "dataset" by the assignment operator "<-".
dataset <- read.csv("dataset.csv")
```

Listing 4 ■ Screening the Imported Data.

```
# Request several descriptive statistics (e.g., mean, median,
# minimum, maximum) of all variables that are part of the data.
summary(dataset)
# Request an overview of the data structure.
str(dataset)
# Request a print of the first six rows of the data set on screen.
head(dataset)
```

Listing 5 ■ Estimating the Overall Effect.

```
# Estimate the overall effect by fitting an intercept-only model.
overall <- rma.mv(y, v, random = list(~ 1 | effectsizeID, ~ 1 | studyID), tdist=
    TRUE, data=dataset)
```

Listing 6 ■ Printing the Results on Screen.

```
# Request a print of the results stored in the object
# ``overall'' in three digits.
summary(overall, digits=3)
```

Output 1 ■ Output of Listings 5 - 6.

```
Multivariate Meta-Analysis Model (k = 100; method: REML)
logLik          Deviance        AIC     BIC     AICc
-73.632         147.264         153.264 161.050 153.517

Variance Components:
                estim   sqrt    nlvls   fixed   factor
sigma^2.1       0.112   0.335   100     no      effectsizeID
sigma^2.2       0.188   0.433   17      no      studyID

Test for Heterogeneity:
Q(df = 99) = 808.848, p-val <.001
Model Results:
estimate        se      tval    pval    ci.lb   ci.ub
0.427           0.118   3.604   <.001   0.192   0.662   ***
---
Signif. Codes: 0 '***' 0.001 '**' '*' 0.05 '.' 0.1 ' ' 1
```

effectsizeID) that were extracted from 17 studies (factor studyID).

- The results of the test for heterogeneity reveal significant variation between all effect sizes in the data set, since the p value is smaller than .001. However, these results are not very informative, as we are interested in within-study variance (level 2) as well as between-study variance (level 3) and not in variance between all effect sizes in the data set.
- The overall effect can be derived from the Model Results. More specifically: estimate = the overall effect size; se = standard error; tval = t value; pval = p value; ci.lb = lower bound of the confidence interval; and ci.ub = upper bound of the confidence interval.

In our example, we can conclude that the overall association between mental health disorders of juveniles and recidivism in juvenile delinquency is 0.427 (expressed in Cohen's d) with a standard error of 0.118. This overall effect is significant ($t(99) = 3.604$, $p < .001$) and the confidence interval is 0.192 to 0.662. According to the criteria formulated by Cohen (1988), stating that d = .2, d = .5, and d = .8 are small, moderate, and large effects respectively, the overall effect of 0.427 can be regarded as small to moderate.

Determining the significance of the heterogeneity in effect sizes

To determine whether the within-study variance (level 2) and between-study variance (level 3) is significant, two separate log-likelihood-ratio tests can be performed. Preferably, these tests are performed one-sided, since variance components can only deviate from zero in a positive direction. In both tests, the null hypothesis states that one of the variance component equals zero, whereas the alternative hypothesis states that the variance component is greater than zero. Performing these tests two-sided would be too conservative (Viechtbauer, 2015, personal communication). In the output of R, p values are by default reported for two-sided tests and since we are performing one-sided log-likelihood-ratio tests, we need to divide the accompanying p values by two.

Heterogeneity of within-study variance (level 2)

Recall from the last output that the variance distributed at the second level of the three-level model was captured in the estimated value of 0.112. For testing the significance of this variance component, we will perform a one-sided log-likelihood-ratio test. In this test, the fit of the original model, in which the variance at the levels 2 and 3 are freely estimated, will be compared to the fit of a model in which only the variance at level 3 is freely estimated and in which the variance at level 2 will be manually fixed to zero. In other words, the fit of the original three-level model will be compared to the fit of a two-level model in which within-study variance is no longer modeled. By doing so, it is possible to determine whether it is at all necessary to account for within-study variance in the meta-analytic model. The null hypothesis in this test states that the within-study variance equals zero ($H_0 : \sigma^2(level2) = 0$), whereas the alternative hypothesis states that the within-study variance is greater than zero ($H_a : \sigma^2(level2) > 0$). If the test results provide support for rejecting the null hypothesis, we can conclude that the fit of the original three-level model is statistically better than the fit of the two-level model, and consequently, that there is significant variability between effect sizes within studies. The significance test can be performed by running the syntax in Listing 7.

This syntax closely resembles the syntax for creating the overall object (see Listing 5), but it has been modified in two respects:

- modelnovar2 = the name of the object in which the results of the rma.mv function will be stored. In our example, we have named this object modelnovar2, since it will contain a model that has no within-study variance at level 2;
- sigma2=c(0,NA) = the argument that is taken by the rma.mv function when the user wants to fix a specific variance component to a user-defined value. The first parameter (0) states that the within-study variance is fixed to zero (i.e., no within-study variance will be modeled), and the second parameter (NA) states that the between-study variance is estimated.

To perform the actual log-likelihood-ratio test, the syntax in Listing 8 needs to be executed.

By calling the anova function, the fit of the two-level model named modelnovar2 will be tested against the fit of the three-level model named overall, which was previously created (see Listing 5). We will now take a look at the output generated by the anova function, which is shown in Output 2.

Output 2 should be interpreted as follows:

- Full represents the three-level model stored in the object overall, whereas Reduced represents the two-level model stored in the object modelnovar2;
- df = degrees of freedom. The reduced model has one degree less than the full model, since within-study variance is not present in the reduced model;
- LRT = likelihood-ratio test. In this column, the value of the test statistic is presented;
- pval = the two-sided p value of the test statistic;
- QE resembles the test for heterogeneity in all effect sizes in the data set, and the value of the test statistic is given in this column. Recall that this test is not very

Listing 7 ■ Building a Two Level-Model without Within-Study Variance.

```
# Build a two-level model without within-study variance.
modelnovar2 <- rma.mv(y, v, random = list(~ 1 | effectsizeID, ~ 1 | studyID),
    sigma2=c(0,NA), tdist=TRUE, data=dataset)
```

Listing 8 ■ Performing a Likelihood-Ratio-Test.

```
# Perform a likelihood-ratio-test to determine the
# significance of the within-study variance.
anova(overall,modelnovar2)
```

informative, as we are interested in both within-study variance (level 2) and between-study variance (level 3) in this three-level meta-analytic example. We are not interested in variance between all effect sizes in the data set.

Given the results, we can conclude that the within-study variance is significant, since the fit of the full model is significantly better than the fit of the reduced model. Simply put, we found significant variability between effect-sizes within studies. Note that the two-sided p value is very small (< .0001) and already smaller than the significance level of .05, so dividing the p value by two does not change this conclusion.

Heterogeneity of between-study variance (level 3)

Determining the significance of the between-study variance proceeds in a similar way. Recall from Output 1 that the variance distributed at the third level of the three-level model was captured in the estimated value of 0.188. We will again perform a one-sided log-likelihood-ratio test, but now, the fit of the original three-level model will be compared to the fit of a model in which only the variance at level 2 is freely estimated and in which the variance at level 3 will be manually fixed to zero. In this last model, between-study variance is not modeled. The null hypothesis in the test states that the between-study variance equals zero ($H_0 : \sigma^2(level3) = 0$), whereas the alternative hypothesis states that the between-study variance is greater than zero ($H_a : \sigma^2(level2) > 0$). If the null hypothesis should be rejected based on the test results, we can conclude that the fit of the original three-level model is statistically better than the fit of the two-level model, and consequently, that there is significant variability between studies. The significance test can be performed by running the syntax in Listing 9.

This syntax has just slightly changed in comparison to the syntax in Listing 7.

- The object is now named `modelnovar3`, since we are determining the significance of the between-study variance at level 3;
- Since we want to fix the between-study variance to zero and freely estimate the within-study variance, we have now typed `sigma2=c(NA,0)`;
- In calling the `anova` function, we have specified that the fit of the two-level model named `modelnovar3` should be tested against the fit of the three-level model named overall.

After running this syntax, output as shown in Output 3 is generated.

Given the results, we can conclude that the between-study variance is significant, since the fit of the full model is significantly better than the fit of the reduced model. Simply put, we found significant variability between studies. Note that the two-sided p value is significant (p < .0001), so dividing this p value by two does not change this conclusion.

In our example, there is significant within-study variance (at level 2) as well as significant between-study variance (at level 3). This implies that there is more variability in effect sizes (within and between studies) than may be expected based on sampling variance alone. Therefore, moderator analyses can be performed in order to examine variables that may explain within- and/or between-study variance. However, before turning to moderator analyses, we will first examine how the total variance is distributed over the three levels of the meta-analytic model.

The distribution of the variance over the three levels of the meta-analytic model

Besides testing the significance of the within-study and between-study variance, it is possible to examine how the total variance is distributed over the three levels of the meta-analytic model. Recall that three different sources of variance are modeled in our meta-analytic model: sampling variance at the first level; within-study variance at the second level; and between-study variance at the third level. To determine how much variance can be attributed to differences between effect sizes within studies (level

Output 2 ■ Output of Listings 7 - 8.

```
         df      AIC     BIC     AICc    logLik  LRT     pval    QE
Full     3       153.264 161.049 153.517 -73.632                 808.8482
Reduced  2       233.156 238.347 233.347 233.281 81.8923 <.0001  808.8482
```

Listing 9 ■ Determining the Significance of Between-Study Variance.

```
# Build a two-level model without between-study variance;
# Perform a likelihood-ratio-test to determine the
# significance of the between-study variance.
modelnovar3 <- rma.mv(y, v, random = list(~ 1 | effectsizeID, ~ 1 | studyID),
    sigma2=c(NA,0), tdist=TRUE, data=dataset)
anova(overall,modelnovar3)
```

2) and to differences between studies (level 3), formulas given by Cheung (2014) can be used. The sampling variance (level 1) cannot be regarded as one fixed value, as this source of variance varies over primary studies. Sampling variance is based on the sample size, and since sample sizes often differ (considerably) from study to study and from effect size to effect size, variation in sampling variance is the consequence. However, it is possible to make an estimate of the sampling variance by using the formula of Cheung (2014, formula 14 on page 2015) and this estimate is also referred to as the typical within-study sampling variance. In Listing 10, the formulas of Cheung are translated into R syntax, with which the distribution of the total variance over the three levels of the meta-analytic model can be determined.

First, we will proceed with an explanation of the syntax in Listing 10.

- In the first eight lines of the syntax, the formula of Cheung (2014, formula 14 on page 2015) is broken down in a number of steps. In each step, a new object is created in which interim results are stored. Eventually, the sampling variance is stored in the object `estimated.sampling.variance`;
- `dataset$v`= variable *v* in object dataset;
- *^2* = squaring an object or variable;
- In creating the objects `I2_1`, `I2_2`, and `I2_3`, each of the three variance components (i.e., sampling variance, within-study variance, and between-study variance, respectively) is divided by the total amount of variance, so that a proportional estimate of each variance component is stored in an object. `overall$sigma2[1]` refers to the amount of within-study variance in the object overall (which was created in Listing 5) and `overall$sigma2[2]` refers to the amount of between-study variance in the object overall.
- In creating the objects `amountvariancelevel1`, `amountvariancelevel2`, and `amountvariancelevel3`, the proportional estimates of the three variance components are multiplied by 100 (%), so that a percentage estimate of each variance component is stored in an object;
- By typing and running the objects `amountvariancelevel1`, `amountvariancelevel2`, and `amountvariancelevel3` seperately, the percentage estimates are printed on screen.

Running this syntax generates the output as presented in Output 4. For ease of interpretation, the last three lines of the syntax in Listing 10 are repeated in Output 4.

From Output 4, we can derive that 6.94 percent of the total variance can be attributed to variance at level 1 (i.e., the typical within-study sampling variance); 34.75 percent of the total variance can be attributed to differences between effect sizes within studies at level 2 (i.e., within-study variance); and 58.30 percent of the total variance can be attributed to differences between studies at level 3 (i.e., between-study variance).

A different approach to heterogeneity

Although performing a significance test is the preferred method for determining whether variance components are significant, it may be wise to examine heterogeneity from a different perspective. A problem that arises in performing log-likelihood-ratio tests is that the test results may not be significant in case the data set is comprised of a rather small number of primary studies and/or effect sizes, even though there is in reality substantial within-study or between-study variance present. In other words, a statistical power problem may be involved. When a research synthesist is presented with non-significant results of log-likelihood ratio tests and consequently decides not to proceed with performing moderator analyses, this may not be the optimal research strategy.

Output 3 ■ Output of Listing 9.

```
         df      AIC     BIC     AICc    logLik  LRT     pval    QE
Full     3       153.264 161.049 153.517 -73.632                 808.8482
Reduced  2       214.066 219.257 214.191 -105.03 62.8024 <.0001  808.8482
```

Listing 10 ■ The Distribution of the Total Variance over the Three Levels.

```
# Determining how the total variance is distributed over the
# three levels of the meta-analytic model;
# Print the results in percentages on screen.
n <- length(dataset$v)
list.inverse.variances <- 1 / (dataset$v)
    sum.inverse.variances <- sum(list.inverse.variances)
squared.sum.inverse.variances <- (sum.inverse.variances) ^ 2
list.inverse.variances.square <- 1 / (dataset$v^2)
sum.inverse.variances.square <-
    sum(list.inverse.variances.square)
numerator <- (n - 1) * sum.inverse.variances
denominator <- squared.sum.inverse.variances -
sum.inverse.variances.square

estimated.sampling.variance <- numerator / denominator

I2_1 <- (estimated.sampling.variance) / (overall$sigma2[1]
    + overall$sigma2[2] + estimated.sampling.variance)

I2_2 <- (overall$sigma2[1]) / (overall$sigma2[1]
    + overall$sigma2[2] + estimated.sampling.variance)

I2_3 <- (overall$sigma2[2]) / (overall$sigma2[1]
    + overall$sigma2[2] + estimated.sampling.variance)

amountvariancelevel1 <- I2_1 * 100
amountvariancelevel2 <- I2_2 * 100
amountvariancelevel3 <- I2_3 * 100

amountvariancelevel1
amountvariancelevel2
amountvariancelevel3
```

Output 4 ■ Output of Listing 10.

```
> amountvariancelevel1
[1] 6.942732

> amountvariancelevel2
[1] 34.75388

> amountvariancelevel3
[1] 58.30339
```

Because of this problem, it may be wise to examine heterogeneity also in a different way by applying the 75% rule as described by Hunter and Schmidt (1990). These scholars state that heterogeneity can be regarded as substantial, if less than 75% of the total amount of variance can be attributed to sampling variance (at level 1). If this is the case, it may be fruitful to examine the potential moderating effect of study and/or effect size characteristics on the overall effect. In our example, approximately 7% of the total amount of variance could be attributed to sampling variance (see Output 4), and based on the rule of Hunter and Schmidt, we can once again conclude that there is substantial variation between effect sizes within studies and/or between studies, making it relevant to perform moderator analyses.

Moderator analyses

Categorical moderators with two categories (i.e., binary or dichotomous predictors)

Because we concluded that there is significant within-study and between-study variance, we are now going to examine whether it is possible to designate variables as moderators of the overall effect. As we use the REstricted Maximum Likelihood estimation method (REML) for estimating the parameters of the meta-analytic model, it is not possible to compare the fit of a model with potential moderating variables to the fit of the model without the potential moderating variables (i.e., performing a log-likelihood-ratio test) (see Hox, 2010; pg. 215). Instead, an omnibus test will be performed to determine whether a (potential) moderating effect of one or more variables included in the model is significant. The null hypothesis in this omnibus test states that all regression coefficients (i.e., betas) are equal to zero ($H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = 0$), and the alternative hypothesis states that at least one of these regression coefficients is not equal to zero. In case an intercept is part of the model (which is the case in our example), it will not be tested in the omnibus test.

In our example, we will first examine the potential moderating effect of publication status of the included primary studies. Recall that two dummy variables regarding publication status are part of the data set: `pstatpub` (coded as 1 = published and 0 = not published) and `pstatnotpub` (coded as 0 = published and 1 = not published). We are going to use both variables in the moderator analysis, but to test whether publication status is a significant moderating variable, we will first extend the meta-analytic model with the variable `pstatpub`. We can test the potential moderating effect of the categorical variable publication status, by running the syntax in Listing 11.

Once again, the syntax in Listing 11 resembles the syntax in Listing 5 that was used for calculating an overall effect, but there are some differences:

- The object in which the results of the moderator analysis are stored has been designated as `notpublished`, because we have chosen the category *not published* (which was coded as 0 in the variable `pstatpub` and coded as 1 in the variable `pstatnotpub`) to be the reference category. Similar to testing categorical predictors in simple regression analysis, one category functions as the reference category and the other categorie(s) are compared against the reference category. From a mere statistical viewpoint, it makes no difference which category is chosen as the reference category;
- `mods =` is the argument that is taken by the `rma.mv` function when the user wants to test the potential moderating influence of a variable. In our example, we are testing whether effect sizes extracted from published studies are significantly different from effect sizes extracted from unpublished studies, and therefore we have added `pstatpub` to the `mods` element by writing `mods = ~pstatpub`. Unpublished studies function as the reference category.

By calling the summary function (see Listing 11), the results as given in Output 5 are presented on screen. The following should be derived from Output 5:

- The results of the `Test for Residual Heterogeneity` show that there is significant unexplained variance left between all effect sizes in the data set ($QE(98) = 702.194$, $p < .001$), after publication status has been added to the meta-analytic model to test its potential moderating effect;
- The results of the omnibus test are presented under `Test of Moderators (coefficient(s) 2)`. The p value is larger than the significant level of .05 and this implies that the regression coefficient of the variable `pstatpub` (the only coefficient that is tested) does not significantly deviate from zero. Therefore, we can conclude that the overall effect is not moderated by the publication status of the included primary studies. The results of the omnibus test can be written as: $F(1, 98) = 1.844$, $p = .178$. Recall that we use the Knapp and Hartung adjustment (Knapp & Hartung, 2003) in our analyses, implying that the omnibus test is based on the F distribution (and not on the normal distribution);
- From the `Model Results`, we can derive the mean effect of the reference category, which is 0.812, and represents the mean effect of the primary studies that have not been published. This mean effect significantly deviates from zero, since $t(98) = 2.656$, $p = .009$. The mean effect of published primary studies is equal to 0.812 + (-

Listing 11 ■ Testing Publication Status as Potential Moderator (Published vs. Unpublished).

```
# Determine the potential moderating effect of publication status;
# Published studies are tested against unpublished studies, so
# unpublished studies serve as the reference category;
# Print the results stored in the object "notpublished"} on screen.
notpublished <- rma.mv(y, v, mods = ~ pstatpub, random = list(~ 1 | effectsizeID, ~
    1 | studyID), tdist=TRUE, data=dataset)
summary(notpublished, digits=3)
```

Output 5 ■ Output of Listing 11.

```
Multivariate Meta-Analysis Model (k = 100; method: REML)
logLik  Deviance        AIC     BIC     AICc
-71.435 142.870         150.870 161.210 151.300

Variance Components:
                estim   sqrt    nlvls   fixed   factor
sigma^2.1       0.113   0.336   100     no      effectsizeID
sigma^2.2       0.171   0.414   17      no      studyID

Test for Residual Heterogeneity:
QE(df = 98) = 702.194, p-val < .001

Test of Moderators (coefficient(s) 2):
QM(df = 1) = 1.844, p-val = 0.178

Model Results:
                estimate        se      tval    pval    ci.lb   ci.ub
intrcpt         0.812           0.306   2.656   0.009   0.205   1.418   **
pstatpub        -0.447          0.329   -1.358  0.178   -1.101  0.206
---
Signif. Codes: 0 '***' 0.001 '**' '*' 0.05 '.' 0.1 ' ' 1
```

0.447) = 0.365 and, as we already learnt from the results of the omnibus test, is not significantly different from the mean effect of unpublished primary studies. The t test statistic used in testing the significance of the regression coefficient of the variable `pstatpub` (-0.447) is not significant ($t(98)$ = -1.358, p = .178) and in line with the result of the omnibus test. Because we are testing only one potential moderating variable in this specific model (i.e., the variable `pstatpub`), the value of the omnibus test (F = 1.844) equals the square of the t-test statistic (-1.358).

Given the results, we can now conclude that the overall association between mental health disorders of juveniles and recidivism in delinquency (d = 0.427) is not moderated by publication status of the included primary studies. If desired, it is possible to examine the significance of the residual within-study and between-study variance, after one or more (potential) moderating variables have been included in the meta-analytic model, by repeating the procedure as described in the sections on heterogeneity of within- and between-study variance, respectively. For now, we are not looking further into the significance of the variance components, since we did not detect a moderating effect of publication status.

It can be of relevance to not only report on the mean effect (including significance and confidence interval) of the reference category, but also on the mean effect (including significance and confidence interval) of the other categories that are tested against the reference category. Above, we manually calculated the mean effect of the other category (i.e., published primary studies in the present example), but for determining the significance and the confidence interval of this mean effect, we need to perform a second analysis. In addition, calculating mean effects

using R is less prone to error than manually calculating mean effects and therefore preferable. For performing this additional analysis, we need to modify the syntax slightly by including the dummy variable `pstatpub` and leaving out the dummy variable `pstatnotpub`. Recall that these two variables are coded in opposite directions, so including `pstatnotpub` in the syntax will give us the mean effect of published studies, which is now the reference category (see Listing 12). Running this syntax generates the output as presented in Output 6.

We can derive from Output 6 that the mean effect of published studies is 0.364 (95% CI: 0.120; 0.609), which is only slightly different from the value we calculated manually (0.365) and this is due to rounding. Note that, in comparison to the results in Output 5, there are no differences in the fit statistics, the estimates of the variance components, and the results of the omnibus test.

Categorical moderators with three categories

Next, we will examine whether the overall association between mental health disorders of juveniles and recidivism in delinquency is moderated by the type of delinquent behavior. We distinguish between three types of delinquency: overt, covert, and general delinquent behavior. Since general delinquent behavior is a non-specific form of delinquency, we wanted this category to be the reference category. This implies that the other two categories (overt and covert delinquent behavior) must be part of the syntax for properly performing the moderator analysis. Recall from section 3 that three mutually exclusive dummy variables representing the three types of delinquency are part of the data set: `typeovert`, `typecovert`, and `typegen`. See Listing 13 for the syntax.

In this syntax, the two variables `typeovert` and `typecovert` have been added. By using the + sign, multiple variables can be added to the `mods` element. Recall that the variable representing the reference category (general delinquency in our example) must not be added to the syntax, otherwise the problem of redundancy will arise. Running the syntax produces the output as presented in Output 7.

From this output, we can derive that:

- There is a moderating effect of type of delinquency, as the results of the omnibus test point towards a significant moderating effect: $F(2, 97) = 7.490$, $p < .001$. This implies that at least one of the regression coefficients of the variables added to the model significantly deviates from zero;
- The mean effect of general delinquency equals 0.470 and this effect significantly deviates from zero: $t(97) = 3.986$, $p < .001$;
- The mean effect of overt delinquency equals $0.470 + (-0.222) = 0.248$. This effect is not significantly lower than the mean effect of general delinquency, as the regression coefficient is not significant: $t(97) = -1.594$, $p = .114$;
- The mean effect of covert delinquency equals 0.470 + (-0.730) = -0.260. This effect is significantly lower than the mean effect of general delinquency, as the regression coefficient is significant: $t(97) = -3.795$, $p < .001$.

Given the results, we can conclude that there is a moderating effect of type of delinquency on the association between mental health disorders and juvenile offender recidivism. For covert delinquency, the association is significantly lower (Cohen's d = -0.260) than for general delinquency (Cohen's d = 0.470). If the research synthesist is interested in testing whether the mean effect of covert delinquency significantly deviates from zero, additional syntax should be written in such a way that the dummy variables *typegen* and *typeovert* are added as potential moderating variables, whereas the dummy variable *typecovert* is left out. In this way, covert delinquency will become the reference category (represented by the intercept), making it possible to determine not only the significance of the mean effect of covert delinquency, but also the confidence interval around this effect. Adding the dummy variables `typegen` and `typecovert` to the syntax (and leaving out `typeovert`), would be necessary if we were to determine the significance of the mean effect of overt delinquency. We could now examine the significance of the residual within-study and between-study variance by repeating the procedure as described in the sections on heterogeneity of within- and between-study variances, respectively. Note that the syntax for creating the objects `modelnovar2` (see Listing 7) and `modelnovar3` (see Listing 9) should be extended with the argument `mods = ~ typeovert + typecovert`, so that the moderator type of delinquency is added to the model.

As a final remark, note that if we were only interested in determining the moderating effect of a discrete variable and not in estimates of the mean effect (including significance and confidence interval) of all the categories of that variable, it would not be necessary to create and test dummy variables. In this case, including that single discrete variable as a moderator in the syntax (i.e., after the `mods ~` element) would suffice. However, it has become rather common to report on the mean effect (as well as significance and confidence interval) of all categories of a discrete potential moderating variable (see, for instance, Assink et al., 2015; Houben et al., 2015; Rapp, Van den Noortgate, Broekaert, & Vanderplasschen, 2014; Van der Hallen, Evers, Brewaeys, Van den Noortgate, & Wagemans, 2015; Van der Stouwe, Asscher, Stams, Dekovic, & Van der Laan, 2014; Weisz et al., 2013).

Output 6 ■ Output of Listing 12.

```
Multivariate Meta-Analysis Model (k = 100; method: REML)
logLik  Deviance        AIC     BIC     AICc
-71.435 142.870         150.870 161.210 151.300

Variance Components:
                estim   sqrt    nlvls   fixed   factor
sigma^2.1       0.113   0.336   100     no      effectsizeID
sigma^2.2       0.171   0.414   17      no      studyID

Test for Residual Heterogeneity:
QE(df = 98) = 702.194, p-val < .001

Test of Moderators (coefficient(s) 2):
QM(df = 1) = 1.844, p-val = 0.178

Model Results:
                estimate        se      tval    pval    ci.lb   ci.ub
intrcpt         0.364   0.123   2.962   0.004   0.120   0.609   **
pstatnotpub     0.447   0.329   1.358   0.178   -0.206  1.101
---
Signif. Codes: 0 '***' 0.001 '**' '*' 0.05 '.' 0.1 ' ' 1
```

Output 7 ■ Output of Listing 13.

```
Multivariate Meta-Analysis Model (k = 100; method: REML)

logLik  Deviance        AIC     BIC     AICc
-66.290 132.581         142.581 155.454 143.240

Variance Components:
                estim   sqrt    nlvls   fixed   factor
sigma^2.1       0.085   0.291   100     no      effectsizeID
sigma^2.2       0.190   0.436   17      no      studyID

Test for Residual Heterogeneity:
QE(df = 97) = 761.162, p-val < .001

Test of Moderators (coefficient(s) 2,3):
QM(df = 2) = 7.490, p-val < .001

Model Results:
                estimate        se      tval    pval    ci.lb   ci.ub
intrcpt         0.470           0.118   3.986   < .001  0.236   0.704   ***
typeovert       -0.222          0.139   -1.594  0.114   -0.498  0.054
typecovert      -0.730          0.192   -3.795  < .001  -1.111  -0.348  ***
---
Signif. Codes: 0 '***' 0.001 '**' '*' 0.05 '.' 0.1 ' ' 1
```

Listing 12 ■ Testing Publication Status as Potential Moderator (Unpublished vs. Published).

```
# Determine the potential moderating effect of publication status;
# Unpublished studies are now tested against published
# studies, so published studies serve as the reference category;
# Print the results stored in the object "published" on screen.
published <- rma.mv(y, v, mods = ~ pstatnotpub, random = list(~ 1 | effectsizeID, ~
    1 | studyID), tdist=TRUE, data=dataset)
summary(published, digits=3)
```

Listing 13 ■ Testing Type of Delinquency as Potential Moderator.

```
# Determine the potential moderating effect of type of delinquency;
# General delinquency is chosen as the reference category;
# Print the results stored in the object "generaldelinquency" on screen.
generaldelinquency <- rma.mv(y, v, mods = ~ typeovert + typecovert, random = list(~
    1 | effectsizeID, ~ 1 | studyID), tdist=TRUE, data=dataset)
summary(generaldelinquency, digits=3)
```

Continuous moderators

In this last example of univariate moderator analyses, we will show how to test a potential continuous moderator. We were interested in examining whether the year in which a primary study was published moderates the overall effect, since changes over time may influence the strength of the association between mental health disorders of juveniles and juvenile offender recidivism. These changes over time may be seen, for instance, in juvenile criminal law or in the way mental health disorders and/or recidivism are operationalized and assessed. We treated publication year as a continuous variable and its potential moderating effect can be tested by running the syntax in Listing 14. Running this syntax produces the output presented in Output 8.

From Output 8, we can derive that:

- Publication year is a significant moderator, as the omnibus test is significant ($F(1, 98) = 5.464$, $p = .021$) and, logically, also the regression coefficient is significant (-0.042; $t(98) = -2.238$, $p = .021$). The regression coefficient is negative, implying that the more recent a primary study has been published, the lower the reported effects in the primary studies;
- The intercept significantly deviates from zero ($t(98) = 4.095$, $p < .001$), but this is not the most important result when testing a continuous moderator. The value of the intercept represents the mean effect of effect sizes extracted from primary studies that have been published in the mean publication year (i.e., when the variable `pyear`, that was centred around its mean, is given the value 0). So, in contrast to the procedure for testing categorical moderators, the intercept cannot be interpreted as the mean effect of a reference category. When testing continuous variables as potential moderators, the regression coefficient (beta) and its significance are in most cases more informative.

In sum, we can conclude that publication year is a significant moderator of the overall association between mental health disorders of juveniles and recidivism in delinquency. As studies have been published more recently (i.e., publication year increases), the strength of the overall association decreases. This significant decrease in effect over time is not indicative for a very robust association between mental health disorders of juveniles and juvenile offender recidivism, and would call for further testing of more specific potential moderating variables. Note that the significance of the within-study and between-study variance can be tested again (see the sections on heterogeneity of within- and between-study variances, respectively), to examine whether there is significant variance left that may be explained by other moderating variables.

Multiple moderator model

In meta-analytic research it is common practice to test the potential moderating effect of multiple variables, such as study, sample, and research design characteristics. As denoted by Hox (2010), many of these variables are often interrelated leading to substantial multicollinearity in the analyses. As a consequence, it is not always straightforward to determine what effects are really relevant and deserve the most attention. In light of this, Hox states that testing multiple moderators in a single model after (potential) moderating effects have been evaluated separately in

Listing 14 ■ Testing Publication Year as Potential Moderator.

```
# Determine the potential moderating effect of publication year;
# Print the results stored in the object "publicationyear" on screen.
publicationyear <- rma.mv(y, v, mods = ~ pyear, random = list(~ 1 | effectsizeID, ~
    1 | studyID), tdist=TRUE, data=dataset)
summary(publicationyear, digits=3)
```

Output 8 ■ Output of Listing 14.

```
Multivariate Meta-Analysis Model (k = 100; method: REML)

logLik  Deviance        AIC     BIC     AICc
-70.282 140.564         148.564 158.904 148.994

Variance Components:
                estim   sqrt    nlvls   fixed   factor
sigma^2.1       0.113   0.336   100     no      effectsizeID
sigma^2.2       0.135   0.367   17      no      studyID

Test for Residual Heterogeneity:
QE(df = 98) = 672.545, p-val < .001

Test of Moderators (coefficient(s) 2):
QM(df = 1) = 5.464, p-val = 0.021

Model Results:
        estimate        se      tval    pval    ci.lb   ci.ub
intrcpt 0.426           0.104   4.095   < .001  0.219   0.632   ***
pyear   -0.042          0.018   -2.238  0.021   -0.078  -0.006  *
---
Signif. Codes: 0 '***' 0.001 '**' '*' 0.05 '.' 0.1 ' ' 1
```

univariate models, is a reasonable strategy. In our final step of the moderator analyses, we follow the approach of Hox and we will examine the unique effect of the variables that were previously identified as significant moderators in the univariate analyses. To do so, we need to extend the meta-analytic model by adding all significant moderating variables simultaneously. In our example, recall that the categorical variable type of delinquency as well as the continuous variable publication year were identified as significant moderators in the univariate analyses (see Outputs 7 and 8, respectively). Therefore, we will extend the meta-analytic model with the variables `pyear`, `typeovert`, and `typecovert`. Since general delinquency was the reference category in testing the variable type of delinquency as a potential moderator, we will not include the dummy variable `typegen` in the syntax. The multiple moderator model can be built by executing the syntax in Listing 15. Running this syntax produces the output as presented in Output 9.

From Output 9, we can derive that:

- At least one of the regression coefficients of the moderators significantly deviates from zero, as the omnibus test shows a significant result ($F(3, 96) = 6.414$, $p < .001$);
- The regression coefficient of publication year (-0.038) significantly deviates from zero, as the t test shows a significant result ($t(96) = -2.077$, $p = .040$);
- The regression coefficient of covert delinquency (-0.709) significantly deviates from zero, as the t test shows a significant result ($t(96) = -3.707$, $p < .001$).

Based on these results, we can conclude that both publication year and the category covert delinquency (versus general delinquency) of the variable type of delinquency have a unique moderating effect on the association between mental health disorders of juveniles and recidivism in delinquency. In other words, we can say that both moderators are robust in the sense that they

Listing 15 ■ Testing Multiple Moderators in a Single Model.

```
# Testing a multiple moderator model in which publication year
# and delinquency type (overt delinquency and covert
# delinquency) have been added as moderators.
multiplemoderator <- rma.mv(y, v, mods = ~ pyear + typeovert + typecovert, random =
    list(~ 1 | effectsizeID, ~ 1 | studyID), tdist=TRUE, data=dataset)
summary(multiplemoderator, digits=3)
```

Output 9 ■ Output of Listing 15.

```
Multivariate Meta-Analysis Model (k = 100; method: REML)

logLik  Deviance         AIC      BIC      AICc
-63.375 126.750          138.750 154.136 139.694

Variance Components:
                estim    sqrt    nlvls    fixed    factor
sigma^2.1       0.085    0.292   100      no       effectsizeID
sigma^2.2       0.149    0.386   17       no       studyID

Test for Residual Heterogeneity:
QE(df = 96) = 609.357, p-val < .001

Test of Moderators (coefficient(s) 2,3,4):
QM(df = 3) = 6.414, p-val < .001

Model Results:
                estimate         se       tval     pval     ci.lb    ci.ub
intrcpt         0.466            0.107    4.346    < .001   0.253    0.678    ***
pyear           -0.038           0.018    -2.077   0.040    -0.074   -0.002   *
typeovert       -0.204           0.139    -1.472   0.144    -0.479   0.071
typecovert      -0.709           0.191    -3.707   < .001   -1.089   -0.330   ***
---
Signif. Codes: 0 '***' 0.001 '**' '*' 0.05 '.' 0.1 ' ' 1
```

are not confounded by the other variable in the model (i.e., covert delinquency (versus general delinquency) is not confounded by publication year and vice versa). This multiple moderator model provides more evidence of true moderating effects of the variables covert delinquency (versus general delinquency) and publication year than the results of the univariate moderator analyses alone. Now that the multiple moderator model is built, it is possible to test the significance of the residual within-study and between-study variance, respectively. Note that, for this purpose, the syntax in Listings 7 and 9 should then be extended with the `mods = ~` argument and all variables that are part of the present multiple moderator model.

Missing data and size of the data set

Although the primary aim of this tutorial is to demonstrate how a multilevel approach can be applied to meta-analytic models in R, we shortly address the problem of missing data in multilevel meta-analytic research. Throughout the years a number of techniques have been developed for assessing whether data is missing in a meta-analytic research project and, if so, how this affects the results. Examples of well-known techniques are the Rosenthal's fail-safe test (1979), Egger's linear regression test (Egger, Davey-Smith, Schneider, & Minder, 1997), the Begg and Mazumdar's Rank Correlation test (Begg & Mazumdar, 1994), and the trim-and-fill method (Duval & Tweedie, 2000a, 2000b). It is good practice for a research synthesist to discuss the extent to which the results were affected by missing data, and to

apply at least one of the available methods for detecting and handling missing data. This also accounts for multilevel meta-analytic research. In scientific literature, there is a considerable and ongoing debate on the appropriateness of the available methods and each method seems to have its own limitations (see, for instance, Egger, Davey-Smith, & Altman, 2001; Nakagawa & Santos, 2012; Nik Idris, 2012; Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin, Schmid, Lau, & Olkin, 2003). Therefore, selecting the most appropriate method for dealing with missing data may not be straightforward. Furthermore, to our knowledge, the available methods have not been evaluated in multi-level meta-analytic research and this makes it even more difficult to select an appropriate method for detecting and handling missing data in a multilevel meta-analytic research project. Evaluating the performance of the available methods in multi-level meta-analysis would be a good direction for future research.

As for the size of the data set used in multilevel meta-analytic research, it is rather difficult to state what the minimum number of studies and effect sizes should be. The statistical power in the analyses increases as the number of studies and effect sizes in the data set increases, but methods for determining the exact power in multilevel meta-analytic models seem not yet available. Further, Viechtbauer (2005) and Van den Noortgate and Onghena (2003) showed that when (restricted) maximum likelihood procedures are used for estimating the parameters in the multilevel meta-analytic model, a smaller number of studies might result in underestimated standard errors and, consequently, an increase in the number of type 1 errors in testing the overall effect size and the moderator effects. In addition, a low number of studies may also lead to the problem of a biased estimate of the between-study variance and the corresponding standard error (see also Van den Noortgate et al., 2013). To be short, larger numbers of studies and effect sizes are to be preferred above smaller numbers, which is not suprising. Future research on the performance and robustness of multilevel meta-analytic models using data sets of different sizes (and types) is needed. All in all, given the difficulties and restrictions of the traditional univariate approach to meta-analysis, the three-level approach in meta-analytic research seems reliable and promising. For further reading on three-level meta-analysis, we refer the reader to Van den Noortgate and Onghena (2003) and Van den Noortgate et al. (2013, 2014).

Conclusion

Applying a multilevel approach to meta-analysis is a strong method for dealing with interdependency of effect sizes, but until today, it is a rather unknown method among scholars and it has not been widely used in meta-analytic research. The main purpose of the present tutorial was to provide an introduction to multilevel modeling in meta-analysis using the rma.mv function of the metafor R package (Viechtbauer, 2015). In specific, we show how the rma.mv function can be called in R syntax, so that a three level structure is applied to a meta-analytic model. In these three-level models, three different variance components are considered: sampling variance at the first level, within-study variance at the second level, and between-study variance at the third level (Cheung, 2014; Hox, 2010; Van den Noortgate, López-López, Marín-Martinez, & Sánchez-Meca, 2013, 2014). In short, this tutorial offers a step-by-step guide for (1) organizing a data file; (2) setting up the R environment; (3) calculating an overall effect; (4) examining heterogeneity of within-study variance and between-study variance; (5) performing categorical and continuous moderator analyses; and (6) examining a multiple moderator model. The statistical approach described in this tutorial has been used in several published meta-analytic reviews (see, for instance, Assink et al., 2015; Gubbels, Van der Stouwe, Spruit, & Stams, 2016; Spruit, Assink, Van Vugt, Van der Put, & Stams, 2016; Spruit, Schalkwijk, Van Vugt, & Stams, 2016; Spruit, Van Vugt, Van der Put, Van der Stouwe, & Stams, 2016). The data file that was used in the present tutorial can be downloaded by the reader from the journal's website.

Authors' note

We thank Prof. Dr. Wim van den Noortgate (University of Leuven) and Dr. Wolfgang Viechtbauer (Maastricht University) for sharing their statistical expertise.

References

Assink, M., Van der Put, C., 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. doi:10.1016/j.cpr.2015.08.002

Begg, C. B. & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101. doi:10.2307/2533446

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

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation

modeling approach. *Psychological Methods*, *19*, 211–229. doi:10.1037/a0032968

Cheung, M. W. L. (2015). *Meta-analysis: A structural equation modeling approach*. New York, NY: John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge Academic.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach (4th ed.)* Thousand Oaks, CA: Sage.

Del Re, A. C. (2015). A practical tutorial on conducting meta-analysis in R. *The Quantitative Methods for Psychology*, *11*(1), 37–50.

Duval, S. & Tweedie, R. (2000a). A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*(449), 89–99. doi:10.1080/01621459.2000.10473905

Duval, S. & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. doi:10.1111/j.0006-341X.2000.00455.x

Egger, M., Davey-Smith, G., & Altman, D. (2001). *Systematic reviews in healthcare*. London: British Medical Journal Books.

Egger, M., Davey-Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634. doi:10.1136/bmj.315.7109.629

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

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*(4), 901–930. doi:10 . 1037 / a0038822

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hunter, J. E. & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings (2nd ed)*. Thousand Oaks, CA: Sage.

Knapp, G. & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693–2710. doi:10.1002/sim.1482

Li, Y., Shi, L., & Roth, D. (1994). The bias of the commonly-used estimate of variance in meta-analysis. *Communications in Statistics - Theory and Methods*, *23*(4), 1063–1085. doi:10.1080/03610929408831305

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

Mullen, B. (1989). *Advanced basic meta-analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nakagawa, S. & Santos, E. S. A. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, *26*(5), 1253–1274. doi:10.1007/s10682-012-9555-5

Nik Idris, N. R. (2012). A comparison of methods to detect publication bias for meta-analysis of continuous data. *Journal of Applied Sciences*, *12*(13), 1413–1417. doi:10.3923/jas.2012.1413.1417

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*(25), 4544–4562. doi:10.1002/sim.2889

R Development Core Team. (2016). R: A language and environment for statistical computing (Version 3.3). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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*(4), 605–618. doi:10.1037/a0036750

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

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *6*, 110–114. doi:10.2307/3002019

Schmidt, F. L. & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings (3rd ed)*. Thousand Oaks, CA: Sage.

Tabachnik, B. G. & Fidell, L. S. (2013). *Using multivariate statistics (6th ed.)* Boston: Allyn and Bacon.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*, 2113–2126. doi:10.1002/sim.1461

Van den Noortgate, W., Lòpez-Lòpez, J. A., Marìn-Martìnez, F., & Sànchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*, 576–594. doi:10.3758/s13428-012-0261-6

Van den Noortgate, W., Lòpez-Lòpez, J. A., Marìn-Martìnez, F., & Sànchez-Meca, J. (2014). Meta-analysis of mul-

tiple outcomes: A multilevel approach. *Behavior Research Methods*, *46*, 1–21. doi:10.3758/s13428-014-0527-2

Van den Noortgate, W. & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, *63*, 765–790. doi:10.1177/0013164403251027

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*(3), 549–573. doi:10.1037/bul0000004

Van der Stouwe, T., Asscher, J. J., Stams, G. J. J. M., Dekovic, M., & Van der Laan, P. H. (2014). The effectiveness of Multisystemic Therapy (MST): A meta-analysis. *Clinical Psychology Review*, *34*(6), 468–481. doi:10.1016/j.cpr.2014.06.006

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*, 261–293. doi:10.3102/10769986030003261

Viechtbauer, W. (2015). Meta-analysis package for R. Retrieved from https://cran.r-project.org/web/packages/metafor/metafor.pdf

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. *JAMA Psychiatry*, *70*, 750–761. doi:10.1001/jamapsychiatry.2013.1176

Wibbelink, C. J. M., Hoeve, M., Stams, G. J. J. M., & Oort, F. J. (2016). *A meta-analysis of the association between mental health disorders and juvenile recidivism*. Manuscript submitted for publication.

Ziegler, S., Koch, A., & Victor, N. (2001). Deficits and remedy of the standard random effects methods in meta-analysis. *Methods of Information in Medicine*, *40*(2), 148–155.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Citation

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*(3), 154–174. doi:10.20982/tqmp.12.3.p154

Received: 14/07/2016 ∼ Accepted: 26/08/2016