

A Practical Tutorial on Conducting Meta-Analysis in R

A. C. Del Re ^a

^a Center for Innovation to Implementation, VA Palo Alto Health Care System, USA

Abstract ■ Meta-analysis is a set of statistical procedures used for providing transparent, objective, and replicable summaries of research findings. This tutorial demonstrates the most common procedures on conducting a meta-analysis using the R statistical software program. It begins with an introduction to meta-analysis along with detailing the preliminary steps involved in completing a research synthesis. Then, a step-by-step tutorial for each of the quantitative components involved for meta-analysis is provided using a fictional set of psychotherapy treatment-control studies as a running example.

Keywords ■ Meta-analysis, R, tutorial, effect sizes

 acdelre@gmail.com

Introduction

Gene Glass (1976) introduced the term meta-analysis to refer to “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings” (p. 3). As with any statistical procedure, meta-analysis has its strengths and limitations (see Table 1), but is now one of the standard tools for providing transparent, objective, and replicable summaries of research findings in the social sciences, medicine, education, and other fields (Hunter & Schmidt, 2004; Hunt, 1997).

This tutorial provides a step-by-step demonstration of the fundamentals for conducting a meta-analysis (summarized in Table 2) in R (R Core Team, 2013). The user should download and install R version 3.1 (or greater) to ensure replicability of each step in this tutorial. Several R packages for meta-analysis will be used (freely available), including `compute.es` (Del Re, 2010) for computing effect sizes and `MAd` (Del Re & Hoyt, 2010) and `metafor` (Viechtbauer, 2010) for aggregating effect sizes, conducting omnibus, meta-regression, and graphics. `MAd` provides a convenience “wrapper” for omnibus and meta-regression functionalities that are available in the `metafor` R package (Viechtbauer, 2010). R is an open-source statistical software program for data manipulation, graphics, and statistical analysis. R can be downloaded freely at <http://www.r-project.org/>.

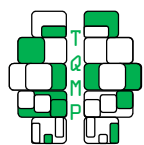
Systematic research strategies

At the start of a meta-analytic endeavor, research questions need to be formulated with precision, as

these questions will affect the entire meta-analytic process. Then, as is usual in any empirical or experimental investigation, inclusion and exclusion criteria must be detailed. This will provide clarity on how the study results may generalize to the population. One of the goals of every meta-analysis is to gather a representative sample of primary studies that meet the study criteria. A systematic research strategy consists of two major steps: (1) defining inclusion and exclusion criteria and (2) selecting studies.

(1) Inclusion and exclusion criteria. Defining study inclusion and exclusion criteria should be based on the study's hypotheses and research questions (see Table 3 for examples). Inclusion/exclusion criteria could potentially bias the study results. Therefore, it is important to be as explicit and thoughtful as possible when defining these criteria.

(2) Study selection. Study selection and the data extraction process are often the most time-consuming steps in conducting a meta-analysis. The study selection process usually follows a particular sequence from the initial search to the coding of effect sizes from the single primary studies. It can be helpful to structure the study selection process based on the 4 steps (study identification, screening, eligibility and inclusion) detailed in the Meta-Analysis Reporting Standards (MARS) guidelines (<http://www.apa.org/pubs/authors/jars.pdf>) or the PRISMA statement (see <http://www.prisma-statement.org/statement.htm>). The above steps should be double coded by two (or more) collaborators to ensure greater objectivity and

**Table 1** ■ Strengths and limitations of meta-analyses.

Strengths.
<ul style="list-style-type: none">- Summarizes a body of research. When a body of research is sufficient enough (publication studies >3), investigation beyond the primary research via meta-analysis is warranted.- Objective and transparent. Meta-analyses are based on well-defined guidelines and a set of procedures, instead of e.g. subjective interpretations.- Robust and replicable. Meta-analytic results (random effects models) will often generalize to the universe of possible study findings in the given area.- Research consolidation. Meta-analysis can serve as catalyst to disentangle relevant from less relevant factors.- Publication bias. Meta analyses allow to estimate publication bias in the report of primary studies.
Limitations
<ul style="list-style-type: none">- Apples and oranges argument. Meta-analysis is criticized for the tendency of analysts to mix incommensurable studies without accounting for differences.- Garbage in, garbage out. Meta-analytic results depend on the methodological quality of the source studies.

precision of the study selection process.

Extracting study-level information and generating reliability statistics

Study characteristics (e.g., average number of sessions in the study) and relevant data to calculate effect sizes should be extracted from each of the included primary studies. Most studies will report more than outcome measure to calculate an effect size. For each study, ensure that the coded effect sizes from a single sample have the same study ID.

The data extracted from the primary studies should be double-coded and checked for the degree to which the two or more coders are in agreement. Double coding is used to determine the degree to which coding errors are present (i.e., reliability), which could subsequently bias the meta-analytic findings. The double coded data should be assessed with interclass correlation procedures for continuous variables (e.g., number of treatment sessions) and Kappa coefficients for categorical variables (e.g. severity of participant

Table 2 ■ General Steps on conducting a Meta-Analysis

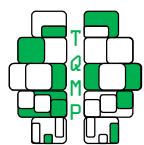
Steps
1. Developing hypotheses/research questions
2. Conducting a systematic search
3. Extracting study-level information and generating reliability statistics <ul style="list-style-type: none">a. Data to calculate effect sizesb. Study-level characteristics (moderators variables)
4. Handling dependent effect sizes
5. Analyzing data <ul style="list-style-type: none">a. Omnibus test(summary effect)b. Heterogeneity testc. Meta-regression
6. Examining diagnostics
7. Reporting findings

distress, coded as “low” or “high”).

Sample data

After the raw dataset is constructed and adequate reliability is obtained for each variable, analyses can begin. For demonstrative purposes, I have simulated (using Monte Carlo simulations) fictional treatment-control psychotherapy data with the following:

- 1- Population (i.e., “true”) effect sizes (based on a normal distribution) of $g = 0.50$, which represents a moderate effect, for outcome one.
- 2- Large population effect sizes of $g = 0.80$ for outcome two.
- 3- An average population sample size of $N = 30$ for both treatment and control groups (average total $N = 60$ for each study)
- 4- Number of sessions as moderator (“dose”; continuous variable)
- 5- Stress level (“stress”; participant baseline severity) in the sample/study dichotomized into “High” and “Low” stress samples

**Table 3** ■ Examples of inclusion/exclusion criteria for psychotherapy meta-analyses

- **Search areas:** Specific journals, data bases (PsycINFO, Medline, Psynindex, Cochrane Library), platforms (EBSCO, OVID), earlier reviews, cross checking references, google scholar, authors contact
- **Written language:** English, German, Mandarin, French
- **Design:** e.g. randomized controlled trials, naturalistic settings, correlational (process-) studies
- **Publication:** Peer review, dissertations, books, unpublished data sets
- **Patient population:** participants inclusion criteria of the primary studies (such as diagnosis), exclusion criteria of primary studies (such as exclusion of substance use disorder), age (e.g. children, adolescents, older adults), participants that were successfully recovered in a prior treatment, number of prior depressive episodes
- **Treatment:** psychotherapy, cognitive behavioral therapy, psychodynamic therapy, medication, time limited / unlimited treatment, relapse prevention, pretreatment training, internet therapy, self-help guideline
- **Outcomes:** All reported outcomes, selected constructs, selected questionnaires, number of participants with clinical significant changes.
- **Comparisons of treatment groups:** Direct comparisons of treatment, indirect comparisons, treatment components, non-bonafide treatments (control/treatment groups that does not intend to be fully therapeutic)
- **Therapists:** Educational level (e.g. masters students, PhD students, licensed therapists), profession (psychologists, psychiatrists, social worker, nurses, general practitioners),
- **Assessment times:** Post assessment, months of follow-up assessments

Therefore, we know that upon completion of this fictional treatment, participants included in the psychotherapy treatment condition will be expected to improve (on average) $\frac{1}{2}$ standard deviation above that of the control group for outcome one and nearly a full standard deviation for outcome two. From the “universe” of possible studies that could have been conducted, 8 studies were randomly selected to use as a running example. We should expect, based on sampling error alone, that there will be variation from the “true” population effects among these studies. That is, no one study will be a perfect representation of the true population effects because each study is a random sample from the entire population of plausible participants (and studies). Details on the meaning of an effect size is provided below.

Getting started

We will begin by first installing the relevant R packages and then loading them into the current R session. At the command prompt (“>”) in the R console, type:

```
library(compute.es) # TO COMPUTE EFFECT SIZES
library(MAd)        # META-ANALYSIS PACKAGE
library(metafor)    # META-ANALYSIS PACKAGE
```

Note that anything following the pound sign (“#”) will be ignored. This is useful for inserting comments in the R code, as is demonstrated above. To follow along with the examples provided in this chapter, first load the following fictional psychotherapy data (which are available when the MAd package is loaded) with

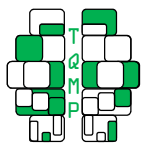
```
data(dat.sim.raw, dat.sim.es)
```

or, run the supplementary R script file (‘tutorial_ma_data.R’) found on the journal’s web site. A description of the variables in these datasets are presented in Tables 4 and 5.

Computing Effect sizes

An effect size (ES) is a value that reflects the magnitude of a relationship (or difference) between two variables. The variance of the ES is used to calculate confidence intervals around the ES and reflects the precision of the ES estimate. The variance is mostly a function of sample size in the study which approaches 0 as the sample size increases to infinity. The inverse of the variance is typically used to calculate study weights, where larger studies are more precise estimates of the “true” population ES and are weighted heavier in the summary (i.e., omnibus) analyses.



**Table 4** ■ Description of dat.sim.raw dataset

Variable	Description
id	Study ID
nT	Treatment group N
nC	Control group N
m1T (m2T)	Treatment group mean for both outcomes
m1C (m2C)	Control group mean for both outcomes
sd1T (sd2T)	Treatment group SD for both outcomes
sd1C (sd2C)	Control group SD for both outcomes
dose	Continuous moderator: Number of therapy sessions
stress	Categorical moderator: Average distress level in sample

For the running example, the ES refers to the strength of the psychotherapy treatment effect. There are several types of ESs, such as standardized mean difference (Cohen's d and Hedges' g) and correlation coefficients (See Table 6). The type of ES used will be dictated by the meta-analytic research question. In our running example, we will compute standardized mean difference ESs, because the outcome is continuous, the predictor is dichotomous (treatment versus control), and means and standard deviations are available. ESs and their variances should be computed for each study to assess the variability (or consistency) of ESs across all included studies and to derive an overall summary effect (omnibus).

The standardized mean difference ES, Cohen's (1988) d , is computed as

$$d = \frac{\bar{X}_1 - \bar{X}_2}{sd_{pooled}} \quad (1)$$

where \bar{X}_1 and \bar{X}_2 are the sample mean scores on the outcome variable at post-treatment in the two groups and sd_{pooled} is the pooled standard deviation of both group means, computed as:

$$sd_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \quad (2)$$

where n_1 and n_2 are the sample sizes in each group and S_1 and S_2 are the standard deviations (SD) in each group. The pooled SD has demonstrated superior statistical properties than other denominator choices

Table 5 ■ Description of dat.sim.es dataset

Variable	Description
id	Study ID
nT	Treatment group N
nC	Control group N
g	Unbiased estimate of d
var.g	Variance of g
pval.g	p-value of g
outcome	Outcome (g1 or g2)
dose	Continuous moderator: Number of therapy sessions
stress	Categorical moderator: Average distress level in sample

(e.g., control group SD at pre-treatment) when calculating d (Hoyt and Del Re, In process). The variance of d is given by:

$$V_d = \frac{2}{\tilde{n}} + \frac{d^2}{2(n_1 + n_2)} \quad (3)$$

where \tilde{n} is the harmonic mean. Cohen's d has been shown to have a slight upward small sample bias which can be corrected by converting d to Hedges' g . To compute g , a correction factor J is computed first:

$$J = 1 - \frac{3}{4 \times df - 1} \quad (4)$$

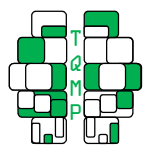
where df is the degrees of freedom (group size minus one). The correction factor J is then used to compute unbiased estimators g and V_g :

$$g = J \cdot d \quad (5)$$

$$V_g = J^2 \cdot V_d \quad (6)$$

For example, to calculate g for outcome one, the `mes` function (which computes ESs from means and SDs) can be run as follows:

```
res1 <- mes(m.1 = m1T, m.2 = m1C,
            sd.1 = sd1T, sd.2 = sd1C,
            n.1 = nT, n.2 = nC, id = id,
            data = dat.sim.raw)
```



where the output from the `mes` function is being assigned (`<-` is the assignment operator) to the object `res1`. `m.1` and `m.2` are the arguments of the function for means of both groups (control and treatment), `sd.1` and `sd.2` are for standard deviations of both groups, and `n.1` and `n.2` are for sample sizes of both groups. See the `compute.es` package documentation for further details.

We will now work with the `dat.sim.es` dataset (see Table 5), which contains ESs (for each outcome) generated from `mes()` using `dat.sim.raw` dataset. This data can be loaded into R (when the `MAd` package is loaded) with `data(dat.sim.es)` at the command prompt. **This dataset is in the 'long' format, where each study provides two rows of data (one for each outcome).** The long data format is ideal for further processing of data that have dependencies among outcomes, as is the case in this sample data. That is, each study reports multiple outcome data from the same sample and these data are not independent. Therefore, prior to deriving summary ES statistics, it is recommended to address dependencies in the data, which will be explained below.

Aggregating dependent effect sizes

Meta-analysis of interventions for a single sample often yield multiple ESs, such as scores reported on two or more outcome measures, each of which provides an estimate of treatment efficacy. The meta-analyst can either choose to either (1) include multiple ESs from the same sample in the meta-analysis, or to (2) aggregate them prior to analysis, so that each independent sample (study) contributes only a single ES.

It is widely recommended to aggregate dependent ESs because including multiple ESs from the same study can result in biased estimates. Further, **ignoring dependencies will result in those studies reporting more outcomes to be weighted heavier in all subsequent analyses.**

There are several options for aggregating ESs. **The most common, but not recommended, procedure is the naïve mean (i.e., simple unweighted mean).** Procedures that account for the correlation among within-study ESs are recommended. These include univariate procedures, such as Borenstein, Hedges, Higgins, and Rothstein (BHHR; 2009), Gleser and Olkin (GO; 1994), and Hedges (2010) procedures and multivariate/multi-level meta-analytic procedures. Based on a large simulation study testing the most common univariate

Table 6 ■ Choice of ES depends on research question

ES	Outcome type	Predictor type
Correlation coefficient	Continuous	Continuous
Mean difference (d or g)	Continuous	Dichotomous
Odds ratio (or Log odds)	Dichotomous	Dichotomous

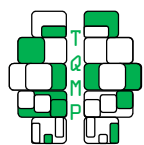
procedures, **BHHR procedure was found to be the least biased and most precise** (Hoyt & Del Re, 2015). This procedure, along with others, have been implemented into the `MAd` package. Aggregating dependent ESs based on the BHHR procedure is demonstrated below:

```
dat.sim.agg <- agg(  
  id = id, es = g, var = var.g,  
  n.1 = nT, n.2 = nC, cor = 0.5,  
  method = "BHHR", data = dat.sim.es  
)
```

In this example above, the output from the aggregation function is being assigned to the object `dat.sim.agg`. The arguments of the `agg` function are `id` (study ID), `es` (ES), `var` (variance of ES), `n.1` and `n.2` (sample size of treatment and control, respectively), `cor` (correlation between ESs, here imputed at $r = 0.5$), `method` (Borenstein, et al. or Gleser and Olkin procedures), and `data` (dataset). The dependent dataset (`dat.sim.es`) is now aggregated into one composite ES per study (dataset `dat.sim.agg`). Note the imputation of $r = 0.5$. This value was chosen because it is a conservative (and typical) starting value for aggregating psychologically-based ESs (e.g., Wampold, Mondin, Moody, et al., 1997). Availability of between-measure correlations within each study are often not available and such starting imputation values are reasonable, although sensitivity analyses with several values (e.g., ranging perhaps from $r = .3$ to $r = .7$, although these values may differ depending on the particular substantive area under investigation) are recommended prior to running omnibus or meta-regression models.

Returning to our data, notice that study-level characteristics (moderators of dose and stress) were not retained after running `agg()`. Therefore, we will **add these moderator variables back into the dataset** and also display a sample of 3 rows of the updated data (`dat.sim.final`), as follows:

```
dat.sim.final <- cbind( dat.sim.agg,  
  dat.sim.raw[, c(12:13)] )  
dat.sim.final[sample(nrow(dat.sim.final), 3), ]
```

which will display a random sample of rows, here being:

	id	es	var	dose	stress
Study 3	0.550	0.052	9	low	
Study 8	0.225	0.048	7	low	
Study 4	1.135	0.048	11	high	

Here **es** is the aggregated ES and **var** is the variance of ES. Note that the appropriate choice of standardized mean difference (*d* or *g*) and its variance to aggregate depends on the aggregation method chosen. Hoyt and Del Re found that using Hedges' *g* (and *Vg*) is preferred when using Borenstein's method (BHR) and Cohen's *d* (and *Vd*) is preferred with Gleser and Olkin's methods (GO1 or GO2).

After within-study dependencies among outcomes have been addressed, the three main goals of meta-analysis can be accomplished. These goals are to (1) compute a summary effect (via omnibus test), (2) calculate the precision or heterogeneity of the summary effect, and (3) if heterogeneity between studies is present, identify study characteristics (i.e., moderators) that may account for some or all of that heterogeneity (via meta-regression).

Estimating summary effect

The summary effect is simply a weighted average of the individual study ES, where each study is weighted by the inverse of the variance (mostly a function of sample size in the study—larger studies are weighted heavier in the omnibus test). There are several modeling approaches for calculating the summary effect and choice of procedure depends on the assumptions made about the distribution of effects.

The fixed-effects approach assumes between-study variance (τ^2) is 0 and differences among ESs are due to sampling error. The fixed-effect model provides a description of the sample of studies included in the meta-analysis and the results are not generalizable beyond the included set of studies. The random-effects approach assumes between-study variance is *not* 0 and those ES differences are due to both sampling error and true ES differences in the population—that is, there is a distribution of “true” population ESs. The random-effect model considers the studies included in the analysis as a sample from a larger universe of studies that could be conducted. The results from random-effects analyses are generalizable beyond the included set of studies and can be used to infer what would likely happen if a

new study were conducted.

The random-effects model is generally preferred because most meta-analyses include studies that are not identical in their methods and/or sample characteristics. Differences in methods and sample characteristics between studies will likely introduce variability among the true ESs and should be modeled accordingly with a random-effects procedure, given by:

$$\theta_i = \mu + v_i^* \quad (7)$$

where θ_i is the true effect for study *i*, which is assumed to be unbiased and normally distributed, μ is the average true ES, and $v_i^* = v_i + \tau^2$, where the variance of the within-study errors v_i is known and the between-study error τ^2 is unknown and estimated from the studies included in the analysis.

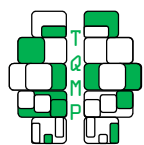
A random-effects omnibus test can be conducted using the **mareg** function in the **MAd** package. This function is a “wrapper”, an implementation of the **rma** function in the **metafor** package (Viechtbauer, 2010). The following code demonstrates the random-effects omnibus test:

```
m0 <- mareg(
  es ~ 1, var = var, method = "REML",
  data = dat.sim.final
)
```

where the output from the omnibus test is saved as **m0** (which stands for “model 0”). The arguments displayed here (see documentation for further arguments) for the **mareg** function are **es** (ES, here the composite *g*), **var** (variance of ES), **method** (REML; restricted maximum likelihood), and **data** (dataset). The **mareg** function works with a formula interface in the first argument. The tilde (~) roughly means “is predicted by”, to the left of ~ is the outcome variable (ES) and to the right is the predictor variable (omnibus or moderators). For omnibus models, 1 is entered to the right of the ~, meaning it is an intercept-only model (model not conditional on any moderator variables). See the documentation in the **MAd** and **metafor** packages for further details. To display the output of the omnibus test, type:

```
summary(m0)
```

which outputs

**Model Results:**

estimate	se	z	ci.l	ci.u	p
0.80	0.13	6.12	0.54	1.05	0

Heterogeneity & Fit:

QE	QEp	QM	QMp
16.23	0.02	37.50	0

where **estimate** is the summary ES for the included studies, **se** is the standard error of ES, **z** is the z-statistic, **ci.l** and **ci.u** is the lower and upper 95% confidence interval (CI), respectively, and **p** is the p-value for the summary statistic. The Heterogeneity statistics refer to the precision of the summary ES and provides statistics on the degree to which there is between-study ES variation. Heterogeneity will be discussed further in the following section.

Interpreting these summary statistics, we find that the overall effect is $g_+ = .80$ (95% CI = 0.54, 1.05), indicating there was a "large" (based on Cohen's interpretive guidelines; 1988) and significant treatment effect at post-testing. In other words, the average end of treatment outcome score for the psychotherapy group was nearly a full standard deviation higher than that of the control group. Therefore, based on the $k = 8$ psychotherapy intervention studies, 67% of treated patients are better off than non-treated controls at post-test. However, based on the Q-statistic (QEp; measure of heterogeneity), there appears to be statistically significant heterogeneity between ESs and further examination of this variability is warranted.

Heterogeneity

As noted above, heterogeneity refers to the inconsistency of ESs across a body of studies. That is, it refers to the dispersion or variability of ESs between studies and is indicative of the lack of precision of the summary ES. For example, suppose the overall summary effect (overall mean ES) for the efficacy of psychotherapy for anxiety disorders is medium (i.e., it is moderately efficacious) but there is considerable variability in this effect across the body of studies. That is, some studies have small (or nil) effects while others have very large effects. In this case, the overall summary effect may not provide the most accurate or important information about this treatment's efficacy.

Why might inconsistencies in treatment effects exist? Perhaps specific study characteristics, such as

treatment duration, are related to this variability. So, suppose that some studies included in the meta-analysis have a short treatment duration and that others have a long treatment duration. If treatment duration had a positive relationship with treatment effect—that is, if ESs increase as the average number of treatment sessions is longer—then the meta-analyst will likely find that the number of treatment sessions accounts for at least some of the variation in ESs. The meta-analyst may then choose to examine other potential moderators of treatment effects, such as treatment intensity, severity of psychiatric diagnosis, and average age in the sample.

There are two potential sources of variability that may explain heterogeneity in a body of studies. The first is the within-study variability due to sampling error. Within-study variability is always present in a meta-analysis because each study uses different samples. The second source of heterogeneity is the between-study variability, which is present when there is true variation among the population ESs.

The most common ways of assessing heterogeneity is with the Q-statistic, τ^2 , and I^2 -statistic. There are other procedures, such as the H^2 -statistic, H -index, and ICC, but they will not be detailed here. The Q-statistic provides information on whether there is statistically significant heterogeneity (i.e., yes or no heterogeneity), whereas the I^2 -statistic provides information on the extent of heterogeneity (e.g., small, medium, large heterogeneity).

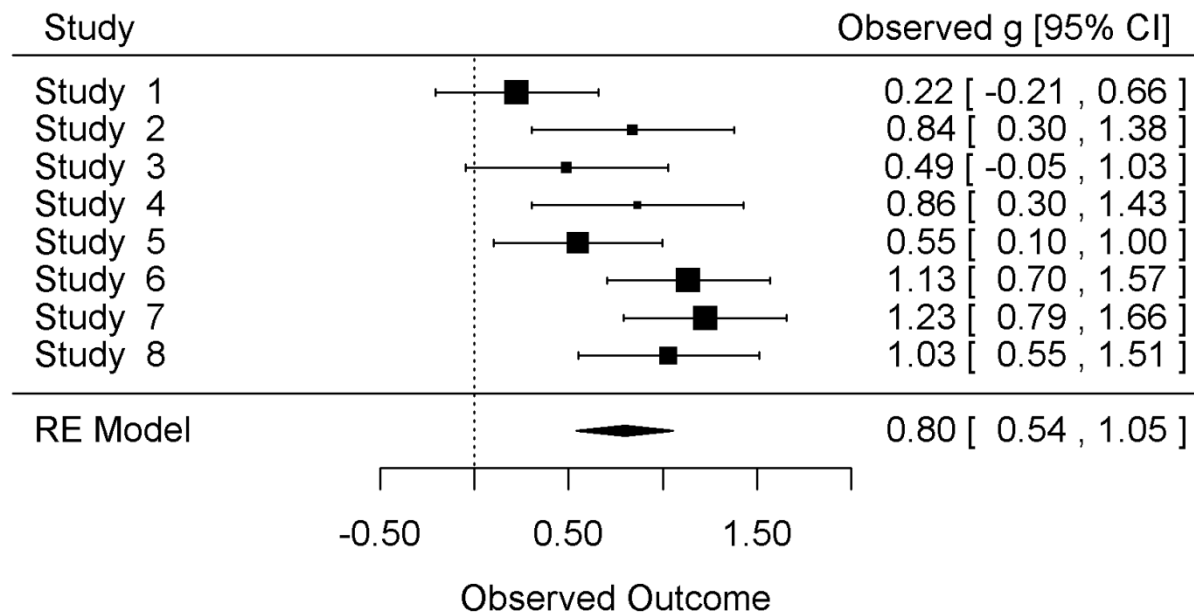
Q is the weighted squared deviations about the summary ES, given by:

$$Q = \sum_i w_i (y_i - \hat{\theta})^2 \quad (8)$$

which has an approximate χ^2 distribution with $k - 1$ degrees of freedom, where k is the number of studies aggregated. Q-values above the critical value result in rejection of the null hypothesis of homogeneity. The drawback of the Q-statistic is that it is underpowered and not sensitive to detecting heterogeneity when K is small (Huedo-Medina, Sánchez-Meca, Marin-Martinez, & Botella, 2006).

I^2 is an index of heterogeneity computed as the percentage of variability in effects sizes that are due to true differences among the studies (Hedges & Olkin, 1985; Raudenbush & Bryk, 2002) and represents the percentage of unexplained variance in the summary ES. The I^2 index assesses not only if there is any between-

Figure 1 ■ Forest plot



study heterogeneity, but also provides the degree to which there is heterogeneity . It is given by

$$I^2 = \left(\frac{Q - df}{Q} \right) 100\%$$

where df is the degrees of freedom ($k-1$). The general interpretation for the I^2 -statistic is that $I^2 = 0\%$ means that all of the heterogeneity is due to sampling error and $I^2 = 100\%$ indicates that all of the total variability is due to true heterogeneity between studies. An I^2 of 25%, 50%, and 75% represents low, medium, and large heterogeneity, respectfully.

Returning to our running example, the heterogeneity output from the omnibus model `m0` will now be clarified. The heterogeneity statistics QE refers to the Q -statistic value, QEp is the p-value for the Q -statistic, QM is the Q -statistic for model fit and QMp is the p-value for QM . The results of the omnibus model indicate there is between-study variation among ESs based on the statistically significant Q -statistic ($QEp = 0.02$). Note that QMp refers to the overall model fit, which equals the p-value for the ES estimate in single moderator models but may change if multi-moderator models are computed (as will be demonstrated below). Although the Q -statistic indicates the presence of heterogeneity between ESs, it does not provide information about the extent of that heterogeneity. To do so, I^2 (and other heterogeneity estimates) and its

uncertainty (confidence intervals) from the omnibus model can be displayed, as follows:

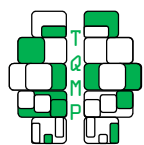
```
confint(m0, digits = 2)
```

which outputs

	estimate	ci.lb	ci.ub
τ^2	0.08	0.00	0.44
τ	0.28	0.03	0.67
$I^2(\%)$	56.14	1.15	88.27
H^2	2.28	1.01	8.52

Here, $I^2 = 56\%$ [1%, 88%], indicating there is a moderate degree of true between-study heterogeneity. However, there is a large degree of uncertainty in this estimate—roughly speaking, we are 95% certain the true value of I^2 is between 1% (all heterogeneity is within-study from sampling error) and 88% (most heterogeneity is due to true between study differences). The large degree of uncertainty in the I^2 estimate is not surprising given low statistical power due to inclusion of a small numbers of studies (Ioannidis, Patsopoulos, & Evangelou, 2007).

The effect and heterogeneity estimates are depicted visually with a forest plot in Figure 1. In this plot, the first column displays the study and the last column displays the details of the ESs and confidence intervals. In the center, each ES is visually displayed (square point) along with their confidence intervals. The size of the point reflects the precision of the ES estimate



(larger studies have larger points). At the bottom, the diamond point represents the summary effect.

Given these results, it appears this treatment effect is not consistent across the body of $k = 8$ studies. Therefore, examining study characteristics (i.e., moderators) that may account for some or all of this heterogeneity is warranted.

Applying moderator models

When ESs are heterogeneous, determining what study characteristics that might account for the dispersion in the summary effect should be considered. However, it is important for meta-analysts to be selective in their analyses and test only those study characteristics for which a strong theoretical case can be made, to avoid capitalizing on chance (Type I error) and identifying spurious moderator variables (Hunter & Schmidt, 2004).

In this fictional dataset, we have identified two moderator variables to examine with mixed-effects models (also called “meta-regression” or “moderator” or “conditional” models). The effects of each moderator variable will be analyzed individually, prior to being analyzed jointly in a multiple moderator model. For example, examining **dose** moderator in a mixed-effects model is given by

$$\theta_j = \gamma_0 + \gamma_1 \times \text{dose} + v_j^* \quad (9)$$

where γ_0 is the expected effect for a study when the moderator is zero, centered at the grand mean, or centered in another way. If a moderator variable accounts for the effects detected, the fixed effect γ_1 will be significantly different than zero (p-values < .05) and the variance, v_j^* , will be reduced. Note that these models have limited statistical power, because the degrees of freedom is based on the number of studies that can be coded for the study characteristic analyzed.

To examine the **dose** moderator (continuous variable) under a mixed-effects model, the following is entered at the R command line:

```
m1 <- mareg(
  es ~ dose, var = var,
  data = dat.sim.final
)
```

```
summary(m1) # DOSE MODERATOR
```

which outputs

Model Results:

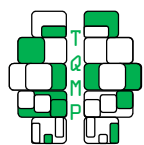
	estimate	se	z	ci.l	ci.u	p
intrcpt	-0.68	0.41	-1.64	-1.49	0.13	0.1
dose	0.15	0.04	3.66	0.07	0.23	0

Heterogeneity & Fit:

QE	QEp	QM	QMp
2.82	0.83	13.40	0

Notice that on the right hand side of the formula, **dose** has now been entered, which can be interpreted as “es predicted by **dose**”. The intercept (**intrcpt**), γ_0 , is -0.68 and, based on the p-value (p ; α set at 0.05), is not statistically significant. Given the lack of significance, this coefficient should not be interpreted. However, if it were significant, it means that when treatment participants have 0 psychotherapy sessions, we expect the average ES to be -0.68. Despite lack of statistical significance, this finding is not particularly meaningful anyway because psychotherapy without a single session is not psychotherapy! Nevertheless, in this case we can roughly interpret the intercept as the average control group ES compared to the treatment group, which, not surprisingly, is similar to the omnibus ES. One possibility for making the intercept value more meaningful, at least in this case (and only for non-dichotomous variables), is to center the moderator variable to a meaningful value. For example, centering the **dose** moderator to the average value across the studies included in the meta-analysis will make the intercept term more meaningful. It will provide information about the average treatment effect for a study with the typical number of psychotherapy sessions (average dosage). This will not be elaborated on any further here but see the R code available on the journal’s web site for examples of how to center moderator variables.

The statistically significant slope coefficient, $\gamma_1 = 0.15$, can be interpreted much like ordinary least squared regression models, where a one unit change in the predictor (here **dose** moderator variable) results in an expected change of γ_1 in the outcome. Given this is a continuous variable, we expect that for each additional psychotherapy session, the ES will increase by 0.15 and we are 95% certain that the true effect lies somewhere between 0.07 and 0.23. Therefore, a participant who completes 10 sessions of this fictional therapy (average number of sessions in this dataset) is expected to



improve by $g = 0.79$ above that of the control group. This can be manually computed at the R command line as follows:

```
y0 <- m1$b[1] # INTERCEPT
y1 <- m1$b[2] # SLOPE
# AVERAGE NUMBER OF SESSIONS = 10
x1 <- mean(dat.sim.final$dose)
# TO CALCULATE EXPECTED ES WITH 10 SESSIONS
y0 + y1 * x1
```

which outputs

```
0.79
```

See Figure 2 (panel a) for a visual representation of the data. The Q-statistic p-value (QE_p) is now nonsignificant, indicating that this moderator accounts for a large proportion of the between-study variance in ESs and reducing τ^2 and I^2 to 0 (which can be displayed by running `confint(m1)` at the command line). However, uncertainty in these heterogeneity estimates are again wide (I^2 range from 0% to 53%), so should be interpreted with caution. **In a non-fictional dataset, it is unlikely that a single moderator will account for all of the between-study heterogeneity.**

The **stress** moderator (dichotomous variable) in a single-moderator mixed-effects model:

```
m2 <- mareg(
  es ~ stress, var = var,
  data = dat.sim.final
)
summary(m2) # SINGLE MODERATOR
```

which outputs

Model Results:

	estimate	se	z	ci.l	ci.u	p
intrcpt	0.97	0.11	9.23	0.76	1.18	0
stressLow	-0.59	0.20	-3.00	-0.97	-0.21	0

Heterogeneity & Fit:

QE	QE _p	QM	QM _p
6.43	0.38	9.04	0

The **stress** variable modeled here is categorical (“high” or “low” values) and the interpretation of slope, γ_1 , is similar but somewhat different than that for continuous variables. Notice that the slope name is **stressLow**. The reason for this is that the reference level for this variable is “high” and the “low” level is being compared to that of the reference level. The high

stress group is represented here by the intercept term and the low stress group is represented by the slope term (γ_1). Therefore, the high stress group improved by $g = 0.97$ (almost a full standard deviation) at post-test compared to the control, whereas the low stress group ES of $g = -0.59$ improved by $g = 0.38$ (i.e., $g_{high} - g_{low} = 0.97 - 0.59 = 0.38$) compared to control. Therefore, therapeutic effects here appear to be moderated by stress levels, such that those with higher stress tend to improve to a greater extent than those in the lower stress group when tested at the end of treatment. See Figure 2 (panel b) for a boxplot representing these findings. **The boxplot is a useful visualization for displaying the distribution of data, where the thick line (usually in the center) represents the median, the box represents 50% of the data, the tails represent the maximum and minimum non-outlying datapoints, and anything outside the tails are outlying data.** Note here there are only two datapoints for the low stress group. In a real situation, one would interpret these findings with great caution, as there are not enough data to be confident in the summary effect for that group. Because these data are for demonstration purposes, we will proceed with further moderator analyses.

Each of the above moderator results are based on single-predictor models. Perhaps these findings are important in isolation but it is crucial to determine the extent to which there may be confounding (i.e., correlation) among these variables. If study characteristics are correlated, results from single moderator models need to be interpreted with caution. Examining the correlation between variables can be accomplish with:

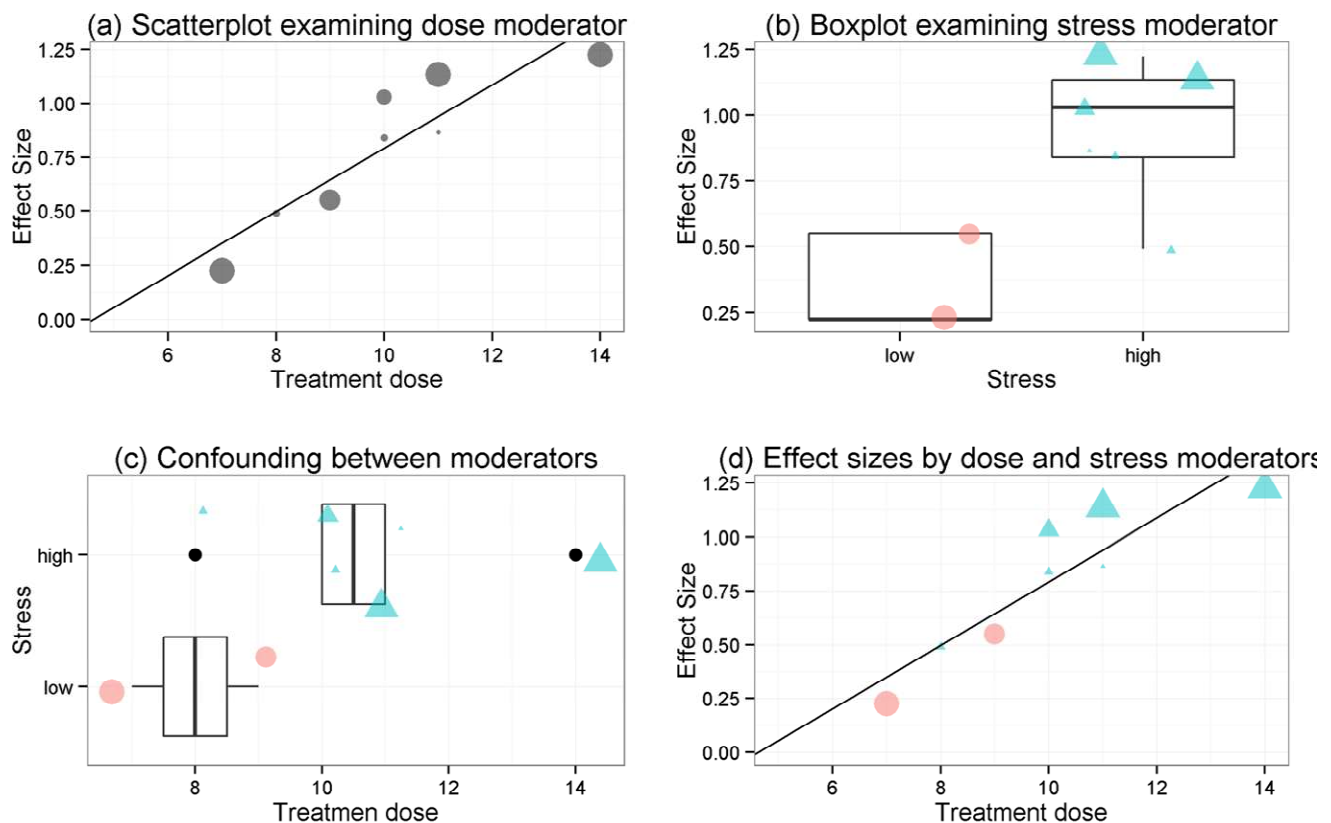
```
with(
  dat.sim.final,
  cor(dose, as.numeric(stress))
)
```

which returns

```
-0.60
```

This indicates that there is a strong (negative) correlation between these variables, although they are not multicollinear (not measuring the same underlying construct). This confounding is displayed as a boxplot (with coordinates flipped) in Figure 2 (panel c), where **stress** is on the y -axis and **dose** on the x -axis. Therefore, it is important to examine these variables in a multi-moderator model to tease out unique variance attributable to each variable while controlling for the

Figure 2 ■ Visualizing continuous and categorical moderator variables



Note. *Hedges' g* effect size (*ES*) is on the y-axes (panels a, b and d) and treatment dose (panel a, c and d) and stress (panel b) are on the x-axes. Each point represents a study and the size of the point represents the study weight (inverse of variance), where larger points are larger sample size studies and are therefore more precise estimates of the population *ES*. Panel a displays the impact of dose on *ES* with the slope coefficient (γ_1) line indicating there is a strong positive relationship. Panel b displays a boxplot of the *ES* distribution for both levels of the stress moderator variable. Panel c displays a boxplot (with coordinates flipped) of the confounding between moderators, such that those in the high stress group have a larger dose than those in the low group. Panel d displays the same information as panel a but has now distinguished the two levels of the stress moderator, where the high stress group is depicted by the blue triangle point and the low stress group depicted by the red circle points.

effect of the other moderator variable.

Confounding among moderator variables.

Ignoring potential confounds among moderator variables may lead to inaccurate conclusions regarding moderator hypotheses (one of the most important conclusions of a meta-analysis, Lipsey, 2003). For example, in the bereavement literature, patient severity has been found to moderate outcome, such that high risk patients ("complicated grievers") have been found to have better outcomes after treatment than "normal" grievers (Currier, Neimeyer, & Berman, 2008). In fact, these findings have sparked debate about the legitimacy of treatment for normal grievers (e.g., should normal grievers even seek treatment?).

However, recent meta-analytic evidence (Hoyt, Del Rey & Larson, 2011) suggests these differences are completely explained by treatment dose (i.e., number of therapy session). They found that studies involving high-risk grievers had longer treatment duration than those studies involving lower risk grievers, which when examined in a multiple-moderator model, the difference in outcome for high-risk and low-risk grievers disappeared (was not statistically significant). Said another way, the amount of treatment is what mattered and not the degree to which the patient was grieving, contrary to findings based on single moderator models.

Returning to the running example, when including both moderators (dose and stress), the formula for

a mixed-effect model yields

$$\theta_j = \gamma_0 + \gamma_1 \times \text{dose} + \gamma_2 \times \text{stress} + v_j^* \quad (10)$$

In this equation, γ_0 is the expected effect for a study when all the moderators are at zero and γ_1 and γ_2 are the expected differences in ES per unit change in each of the moderators, while holding the other moderator constant. If γ_1 remains statistically significant in this model, it can be inferred that the treatment dosage (dose) is a robust moderator in the sense that it is not confounded with the other moderator.

The code and output for this model yields a strikingly similar effect as

```
m3 <- mareg(
  es ~ dose + stress, var = var,
  data = dat.sim.final
)
```

```
summary(m3) # MULTIPLE MODERATOR
```

which outputs

Model Results:

	estimate	se	z	ci.l	ci.u	p
intrcpt	-0.26	0.57	-0.45	-1.38	0.87	0.70
dose	0.11	0.05	2.18	0.01	0.21	0.03
stresslow	-0.26	0.24	-1.07	-0.73	0.22	0.29

Heterogeneity & Fit:

QE	QEp	QM	QMp
1.69	0.89	14.54	0

the findings from the grief meta-analysis (which was intentional for demonstrative purposes). Here we find that with both moderator variables in the model, **dose** remains statistically significant but **stress** becomes nonsignificant. This is displayed visually in Figure 2 (panel d). Therefore, as in the grief meta-analytic findings, when controlling for treatment dose, stress levels are not relevant. And, expectedly (given the single predictor heterogeneity estimates), the p-value for the overall model fit is significant (QMp) and the Q-statistic is nonsignificant (although uncertainty in τ^2 and I^2 remains large, e.g., I^2 range from 0% to 50%). This demonstrates the importance of taking caution when interpreting single moderator models without

first examining potential confounding between moderator variables.

Sensitivity analyses and diagnostics

Given space constraints, this section will not be afforded the attention it deserves but the reader is referred to Viechtbaur and Cheung's (2010) study on diagnostics for meta-analysis. There are several diagnostic tools available for meta-analysis, including tools for publication bias (e.g., funnel plot, trim and fill, fail-safe N) and tests to determine the impact/influence of a study on omnibus and mixed-effects outcome (e.g., Cook's distance, externally standardized residuals, radial plots). The meta-analyst should *always* examine for publication bias and influential studies.

An examination for potential publication bias ("file-drawer problem") is generally the first recommended diagnostic test and will be the only one examined in this section. Publication bias refers to the possibility that studies showing a statistically significant effect are more likely to be published than studies with null results which could bias the summary effect. The funnel plot is a useful visualization to examine for publication bias. It is a scatter plot of the ES (on the x-axis) and a measure of study precision (generally the standard error of ES on the y-axis). Asymmetry in the funnel may be indicative of publication bias but in some cases this plot can be fairly subjective. Creating a funnel plot can be achieved with:

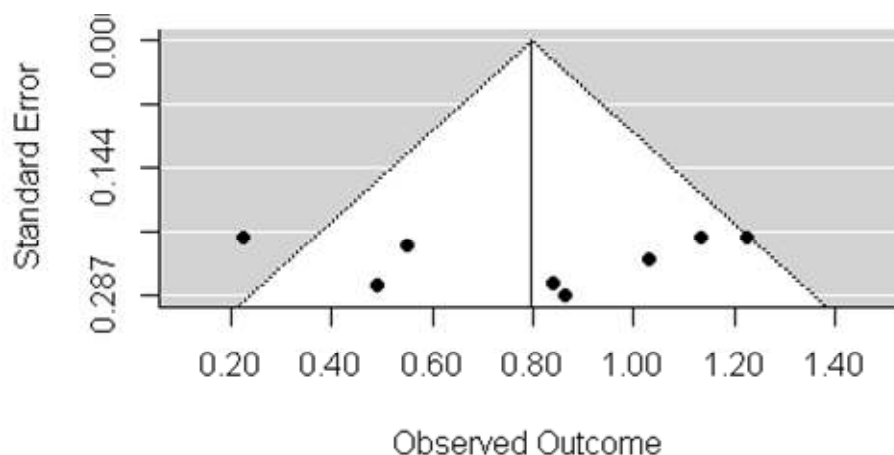
```
# EXAMINE PUBLICATION BIAS VISUALLY
funnel(m0)
```

Based on the plot (Figure 3) there is no visual indication of publication bias. There are several additional publication bias and diagnostic procedures demonstrated in the R file on the journal's web site and the reader is referred there for details on these procedures.

Reporting findings

As in other empirical investigations, the quality of a meta-analytic study is only as good as its clarity, transparency, and reproducibility. Precision in documenting each of the many steps involved in conducting a meta-analysis can be challenging. It can be useful to keep a diary about the various decisions – sometimes even the small ones – that were made at each step. The MARS guidelines and the PRISMA statement offer a standard way to check the quality of reports, and to ensure that readers have the

Figure 3 ■ Funnel plot.



information necessary to evaluate the quality of a meta-analysis.

Conclusion

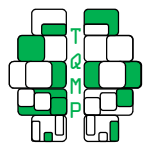
Meta-analysis is one of the standard tools for providing transparent, objective, and replicable summaries of research findings. In this tutorial, many of the basic steps involved in conducting a meta-analysis were detailed and demonstrated using R packages that are freely available for meta-analysis. Interested readers can reproduce each of the steps demonstrated here by running the supplementary R script file found on the journal's web site.

Author's note

I would like to thank Dr. Christoph Flückiger for his contributions to an earlier version of this paper.

References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum.
- Cooper, H. M., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation Publications.
- Currier, J. M., Neimeyer, R. A., & Berman, J. S. (2008). The effectiveness of psychotherapeutic interventions for bereaved persons: a comprehensive quantitative review. *Psychological Bulletin*, 134(5), 648.
- Del Re, A. C. (2010). *compute.es: Compute Effect Sizes*. Madison, WI.
- Del Re, A. C., & Hoyt, W. T. (2010). *MAd: Meta-analysis with mean differences*. Madison, WI.
- Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed-and random-effects methods. *Psychological Methods*, 6(2), 161.
- Fox, J., Ash, M., Boye, T., Calza, S., Chang, A., Grosjean, P., & Wolf, P. (2007). Rcmdr: R commander. *R package version*, 1-3.
- Hedges L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press Orlando, FL.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486.
- Hoyt, W. T. & Del Re, A. C. (2015). Comparison of methods for aggregating dependent effect sizes in meta-analysis. Manuscript under review
- Hoyt W. T., Del Re, A. C., & Larson, D. G. (2011, June). A new look at the evidence: Grief counseling is effective. Paper presented at the *9th International Conference on Grief and Bereavement in Contemporary Society*, Miami, FL.
- Huedo-Medina, T. B., Sánchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or *I* index? *Psychological Methods*, 11(2), 193.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage Publications, Inc.
- Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. Russell Sage Foundation. Retrieved from <http://books.google.com/books?hl=en&lr=&id=0r5hxjbT6pQC&oi=fnd&pg=PR11&dq=hunt+1997+meta+analysis&ots=5sVieCPjS-&sig=IMynBOC7-JTxJ4YKQ3vzgvSe2OE>



- Ioannidis, J. P., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *British Medical Journal*, 335(7626), 914.
- Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The Annals of the American Academy of Political and Social Science*, 587(1), 69-81.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage Publications, Inc.
- R Core Team (2013). *R: a language and environment for statistical computing. 2013. R Foundation for Statistical Computing, Vienna, Austria.*
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1-48.
- Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1(2), 112-125.
- Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. N. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "all must have prizes". *Psychological bulletin*, 122(3), 203.

Citation

- Del Re, A. C. (2015). A Practical Tutorial on Conducting Meta-Analysis in R. *The Quantitative Methods for Psychology*, 11(1), 37-50.

Copyright © 2015 Del Re. This is an open-access article distributed under the terms of the *Creative Commons Attribution License (CC BY)*. The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 13/12/14 ~ Accepted: 16/01/15