19
Meta-analysis

S. Natasha Beretvas

Meta-analysis entails a set of analytical techniques designed to synthesize findings from studies investigating similar research questions. While meta-analysis includes narrative integration of results, the current chapter will focus only on quantitative meta-analysis. Meta-analysis permits summary of studies' results and is designed for scenarios in which the primary studies' raw data are not available. The meta-analytic process involves summarizing the results of each study using an effect size (ES), calculating an overall average across studies of the resulting ESs, and exploring study- and sample-related sources of possible heterogeneity in the ESs. The overall average ES provides a single best estimate of the overall effect of interest to the meta-analyst. Meta-analysis can be used to explore possible differences in ESs as a function of study and sample characteristics. In the seminal article in which the term *meta-analysis* was coined, Smith and Glass (1977) used meta-analysis to summarize results from studies that had assessed the effectiveness of psychotherapy. Thus, treatment effectiveness results provided the first type of ES to be synthesized using meta-analysis. Since the 1970s, the field of meta-analysis has grown to include methods for conducting the synthesis of other types of ESs including correlations, transformations of odds-ratios, validity coefficients, reliability coefficients, and so forth.

Many textbooks provide detailed descriptions of the meta-analytic process. Texts by Lipsey and Wilson (2001), Rosenthal (1991), Card (2012), and Borenstein, Hedges, Higgins, and Rothstein (2009) provide excellent introductions to meta-analysis. Hunter and Schmidt's (1990) textbook provides the seminal resource for meta-analysts interested in correcting ESs for artifacts (see Desideratum 11). Books by Cooper, Hedges, and Valentine (2009) and Hedges and Olkin (1985) are recommended for readers with more technical expertise. Meta-analysts interested in a text devoted to description of ways to assess and correct for publication bias should refer to Rothstein, Sutton, and Borenstein (2005). Desiderata for studies that involve use of meta-analysis are contained in Table 19.1 and thereafter they are discussed in further detail.

1. Theoretical Framework and Narrative Synthesis

As with any manuscript, a summary of past research must justify the selection of the study's research question. Similarly, a meta-analysis must be prefaced by a narrative synthesis summarizing results found in previous studies that are to be integrated in the meta-analysis. The narrative synthesis must clarify the specific research question associated with the effect size (ES) that is being synthesized.

Table 19.1 Desiderata for Meta-analysis.

Desideratum	*Manuscript Section(s)*[*]
1. A theoretical framework is provided that supports the investigation of the effect size (ES) of interest and includes a narrative synthesis of previous findings.	I
2. Type of ES of interest in the study is specifically detailed (e.g., correlation, standardized mean difference).	I, M
3. Databases searched and keywords used to find relevant studies are listed, as well as criteria for deciding whether to include a study in the meta-analysis.	M
4. Formulae used to calculate ESs are provided or referenced, and any transformations used (e.g., to normalize or stabilize ES sampling distributions) are made explicit.	M
5. The coding that is used to categorize study and sample descriptors is provided.	M
6. Estimates are provided that describe the interrater reliability of the information coded in each study.	M, R
7. If study quality is assessed, a description is provided detailing how it is assessed and how study quality is incorporated into the meta-analysis.	M
8. For weighted analyses, the type of weights used is provided.	M
9. Methods used to handle within-study ES dependence (e.g., multiple ESs per study) are described.	M
10. Methods used to access, assess, and handle missing data are detailed.	M
11. If relevant, the method used to correct for artifacts is described.	M, R
12. Homogeneity of ESs is assessed.	M, R
13. Statistics describing the resulting meta-analytic dataset that was gathered and including pooled estimates of the effect size of interest are provided along with associated standard errors (and/or confidence intervals).	R
14. Inferential statistics describing the relation between the study and sample descriptors and the effect size are presented.	R
15. Interpretation is offered describing the practical significance of the ES magnitude and direction and the relation between moderators and the ES.	D

[*] I = Introduction, M = Method, R = Results, D = Discussion.

The narrative synthesis summarizes in words what previous research has found in terms of the patterns of results relevant to the ES of interest. While many studies might investigate the same basic research question, the studies can be distinguished by various sample and study composition descriptors. Examples of descriptors include demographic variables such as gender, ethnicity and age, and characteristics of the study's design such as the type and duration of an intervention, the outcome measure used, the research context, and the experimental design used. The review of previous literature should clarify and identify the importance and relevance of these descriptors to the ES of interest. This then lays the groundwork for investigation of relations between these descriptors (termed *moderators*) and the ES in the ensuing meta-analysis.

2. Effect Size

The fundamental unit of any meta-analysis is the effect size (ES). An ES provides a parsimonious descriptor containing information about the direction and magnitude of the results of a study. The most commonly used meta-analytic ESs include the standardized mean difference, the correlation (representing the relation between two variables), and the odds ratio. A meta-analytic ES describes the relation between a pair of variables. Operationalization of each of the two variables should be clarified and justified. For example, if student achievement is one of the pair of variables of interest in a meta-analysis then the sorts of test scores that qualify as student achievement should be clarified. Description of the research question of interest in the meta-analysis should clarify the ES being investigated both in terms of the statistical type as well as the operationalization of each of the relevant two variables.

3. Study Inclusion Criteria

The Introduction (see Desiderata 1 and 2) should have clarified the components necessary for deciding to include a study's results in the meta-analysis. A section in the Methods section of a meta-analysis must detail how the relevant studies and results were found. The databases (e.g., PsycInfo, ERIC) that were searched, the types of studies (e.g., peer-reviewed publications, dissertations, conference presentations) and the keywords used must be identified. Any additional means used for finding relevant studies that were not initially identified should also be described (e.g., using the References section in studies that had been identified in the database search, contacting study authors, etc.). In addition to emphasizing the acceptable operationalizations of the constructs relevant to the ES, the population of interest should be described. For example, a researcher might solely be interested in an ES for adults and thus data would be excluded from any study that had investigated the relevant variables for adolescent respondents.

Meta-analysts must also decide on the types of study designs that qualify for inclusion. Some meta-analysts include only results from studies employing purely experimental designs while others also include quasi-experimental studies' results. Some studies necessitate the use of single-subject designs for which there is still controversy in terms of how to meta-analytically synthesize the results. If a more general inclusion strategy is used, then meta-analysts should code the relevant design features and summarize descriptively or inferentially the potential differences in resulting ESs (see Desiderata 13 and 14).

4. Calculation of Effect Sizes

The (statistical) type of ES being synthesized should have been clarified in the Introduction (see Desideratum 2). Results reported in the primary studies being synthesized are not all in the same format. For example, a meta-analyst might be interested in synthesizing a treatment's effectiveness using a summary of standardized mean differences across studies. Some studies might provide the treatment and control groups' means and standard deviations for the relevant outcome. Other studies might instead provide the results of an independent samples t-test comparing the treatment and control groups on the outcome score. Results in both formats can be converted into a standardized mean difference ES metric. Authors should clarify any conversion formulas they use to convert studies' results into a common ES metric.

In addition, some estimators of the most commonly used ES (the standardized mean difference) have been found to be biased. There are a number of ways that this ES is calculated (including, most commonly, Cohen's d, Glass's Δ, and Hedges's g). The meta-analyst must clarify and justify which estimate of the standardized mean difference is being used.

Sampling distributions of most of the typical untransformed ESs (e.g., standardized mean difference, correlation, odds ratio) have been found to be non-normal. One of the purposes of quantitative meta-analysis is to use statistical tests of the ES and its relation with sample and study descriptors. Thus, it is important to use the transformations that normalize (and stabilize the variances) of the sampling distributions of these ESs. Meta-analysts should detail the formulas that are used to transform the resulting ESs estimates for ensuing statistical analyses.

5. Coding of Study and Sample Descriptors

A host of variables typically distinguish the studies and samples being synthesized and might be related to the resulting ESs. Sources of the possible heterogeneity in ESs across studies can and should be explored using these variables. When gathering primary study data to be used for calculating the ESs, meta-analysts should also gather information associated with the samples in each study. Sample size is an essential variable that must be coded as it provides information about the precision of each study's ES and can be used as a weight in resulting ES analyses (see Desideratum 8). Demographic information (such as age, gender, and ethnicity composition of the sample) can also be coded and used in the meta-analysis. Characteristics of each of the two variables whose relation is being synthesized should also be coded and captured. For example, in a study summarizing a family-based treatment's effectiveness in reducing internalizing disorders, the meta-analyst might have multiple constructs such as depression and anxiety that qualify as internalizing disorders. Each type of outcome could be coded to explore possible differences in the treatment's effectiveness for the more specific kinds of internalizing disorders. This can lead to multiple ES estimates being gathered per study and thus some dependence that must be handled (see Desideratum 9). There might also be characteristics of the implementation of the treatment that distinguish the primary studies and define the resulting ESs. In the current internalizing disorders example, interventions might be designed to involve both parents and children or they might be designed only for parents. Thus, categories distinguishing interventions could also be coded and collected. In addition, and specifically for intervention effectiveness meta-analyses, some studies might report results for more than one intervention. As with a study reporting multiple outcomes, the dependence resulting from multiple ESs per study needs to be appropriately handled (see Desideratum 9). Last, for a meta-analysis of intervention effects, not all studies might compare all interventions of interest. Network meta-analysis procedures can be used to synthesize results from studies comparing different sets of interventions' effects with each other.

Facets of a study's design can also be gathered and included in the meta-analysis (as described in Desideratum 14). As mentioned in Desideratum 3, a study's design should be coded as it can later be used to explore potential differences in ESs resulting from differing experimental designs. Some meta-analysts code "study quality" and evaluate its relation to the ES values. Some meta-analysts correct their ESs for artifacts to match what the ESs would be for a perfect study that used an infinitely large sample with access to perfectly reliable and valid test scores. If interested in correcting for artifacts, the meta-analyst would gather relevant information including, for example, the reliability of scores on the measures of interest (see Desideratum 11). Additional selection of study and sample descriptors should be founded in the meta-analysts' research questions in terms of what they hypothesize might explain variability in ESs.

Values for some of the descriptors might differ for samples within a study. Group sample size in a meta-analysis of a treatment's effectiveness (i.e., using the standardized mean difference ES to summarize the difference in means between a treatment and control group) provides a simple example of a sample-level descriptor. Other descriptors might only vary across studies (e.g., whether the population being assessed was college students). The coder must clarify the distinction between such sample-level descriptors and study-level descriptors that differ across, but not within, studies. This information is essential to inform selection of the analytic technique that best matches the data's structure and for addressing how to mean-center moderators in meta-regression (moderator) analyses.

One last piece of information about coding must also be provided. Unfortunately, the information sought by meta-analysts is not always presented in the primary studies. It is important for meta-analysts to clarify how they attempted to gather moderator variable values that are missing as well as to detail the methods used to handle the missingness (see Desideratum 10).

6. Interrater Reliability

Given the amount of information that needs to be gathered and coded in a meta-analysis, it is typical to involve at least a couple of researchers as coders. It is thus important to provide a description of the reliability of the coding that was conducted. If data indicate that coding is not reliable, further coding training should be conducted and consensus about each study's codes must be reached. At the very least, the average (median) percent agreement for each variable should be reported in the meta-analysis. Use of kappa, weighted kappa, or the intraclass correlation to provide additional measures of interrater agreement is also encouraged (see Orwin & Vevea, 2009, for additional details; see also Chapter 10, this volume). While it would be optimal for at least two coders to code every study in the meta-analysis, that sometimes is not feasible. If this is the case, then at least a reasonable proportion of studies should be coded by at least two raters with sufficient justification provided for not having two raters code all studies. Given a lack of complete agreement in the coding that is done by the two raters, the meta-analyst must describe how differences were resolved and consensus reached through discussion and possible re-specification of codes used.

7. Study Quality

Since the introduction of the term *meta-analysis* in the 1970s (Glass, 1976), researchers have argued about how to handle differences in research designs' quality when synthesizing studies' results. Researchers agree that, at the outset, meta-analysts must select and justify a research design quality criterion for study inclusion. In addition, meta-analysts are encouraged to gather and code information (see Desideratum 5) on a study's design that might differentiate studies' ES results.

All sorts of factors might impact the quality of a study's design and thus also affect the ES results. Those factors include group selection and assignment, experimenter expectations (e.g., whether a study is blinded), psychometric properties of measures, and many more. It is up to the researcher to select the pool of possible design quality variables of relevance to the meta-analysis. Meta-analysts can use the resulting variables descriptively or use them as moderating variables in ensuing analyses (see Desiderata 13 and 14).

8. Weights

As with any consistent estimator, the precision of an ES estimate is greater when it is based on larger sample sizes. Thus, when pooling ES estimates, meta-analysts typically weight ESs by some function of their associated sample sizes (see Desideratum 13). When testing meta-regression models (see Desideratum 14) designed to explore the variability in ESs using study and sample descriptors as moderators, meta-analysts frequently estimate models involving these same *N*-based weights. In either scenario, more weight is assigned to estimates based on larger sample sizes. The most commonly used weights are either the inverse of *N* or the inverse of the variance of the ES of interest (which will also be a function of *N*). The weight entailing the inverse of the conditional variance results in the most efficient pooled estimate of the population ES and thus is recommended here. However, the meta-analyst should clarify the function of *N* that is being used as the weight.

9. Handling Dependent ESs

Studies can frequently contribute multiple ESs to a meta-analysis. These multiple ESs can be considered dependent if they are based on the same sample. For example, in meta-analyses designed to assess intervention effectiveness (i.e., comparing two groups on an outcome), a study can provide results from comparing the two groups on each of multiple, related outcomes. Given that sufficient data are provided in the study for each outcome that corresponds to the construct of meta-analytic focus (e.g., depression and anxiety might both qualify as internalizing outcomes), an ES can be calculated. The resulting two standardized mean difference ESs are assumed dependent because the ESs describe a common sample. Equally important is that the ESs are based on measures that are themselves correlated (e.g., depression and anxiety).

Alternatively, in a meta-analysis of the correlation between two variables, multiple dependent ESs would result from a study that provided correlation estimates between pairs of variables both of which matched the constructs of interest. This study would qualify as a *multiple-endpoint* study. For example, the meta-analyst might be interested in the correlation between internalizing disorders and academic achievement. If a study reports the correlation between, say, depression and SAT scores and the correlation between anxiety and SAT scores, then both correlations could be used to calculate ESs for later analysis. The dependence would again originate in the use of a common sample for estimation of the two ESs.

Another example of the source of possible dependence commonly found in meta-analyses of intervention research might originate in a study reporting results from comparing three groups on an outcome. This study would be an example of a *multiple-treatment* study. For example, a meta-analyst might be interested in summarizing the effectiveness of parental involvement interventions for improving internalizing disorders. A primary study might evaluate the internalizing disorders of three groups, two of which involve differing implementations of a parental involvement treatment and a control group. Two effect sizes could be calculated with one comparing the internalizing disorder scores of the first intervention group with the control group. The second ES would describe the difference in internalizing disorders between the second intervention group and control group. Given the involvement of the same control group in the calculation of the two ESs, the ESs would be considered dependent. Other possible dependencies might be encountered between pairs of effect sizes within a study that should also be handled appropriately.

Meta-analysts have a choice of methods they can use to handle effect size dependence. Some researchers choose to ignore the dependence which will negatively impact the validity of the associated statistical conclusions. Other meta-analysts might choose a single effect size to represent each study. For example, this "best" ES might be based on the measure with the best psychometric functioning in each study. Still others might calculate a weighted or simple average of each study's multiple ESs and use the result as the single ES for each study. While use of a single ES per study (selected via aggregation or deletion of the study's multiple ESs) does result in an analysis of independent ESs, it overly reduces the available database and thus possible information. In addition, this reduces ensuing statistical power. It also unnecessarily reduces the possible heterogeneity in the ESs.

Another option available for handling dependent ESs involves modeling the multivariate nature of the dataset. Several options are available with use of generalized least squares (GLS) estimation procedures being the most commonly used method. The primary problem with the use of multivariate modeling to handle possible dependencies is that additional data must be gathered from the primary studies. For example, to use GLS for synthesizing results from multiple-endpoint studies, meta-analysts must use values for the correlation among scores on the multiple endpoints. However, it is sometimes possible to impute reasonable values for this correlation and, despite their complexity, GLS methods have been found to work well for handling meta-analytic dependence in some scenarios. Two more recent methods suggested for handling within-study dependence include the use of robust variance estimation (Hedges, Tipton, & Johnson, 2010) and the multilevel meta-analysis model (van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). Regardless of the method used, the meta-analyst must note the types of dependence that they encountered in their dataset. They must also describe and justify their choice of method used to handle this dependence.

10. Methods for Handling Missing Data

As with most social science datasets, analysis of meta-analytic datasets is also hampered by missingness. This can result from primary studies not reporting sufficient statistical information permitting calculation of an effect size. Alternatively, primary studies might not have gathered or not reported all information of interest to the meta-analyst. For example, a meta-analyst might be interested in explaining heterogeneity in an ES using a variable representing the percent of participants who were female. Not every study will necessarily report the percent of participants who are female. Meta-analysts need to detail and justify how they handled missing data.

As with primary study analyses, there are a host of options that meta-analysts can use to handle missingness. There are similar caveats associated with these techniques when used in a meta-analytic context. For example, use of listwise or pairwise deletion still requires the assumption that data be missing completely at random and frequently results in large reductions in data available for a meta-analysis. These methods are not strongly recommended for use with meta-analytic data. Use of single-value imputation is not uncommon in meta-analysis (e.g., using a mean of reported values' information, or using a value that is reasonable based on patterns of values reported in other studies with similar participants). Single-value imputation can be recommended although its use inappropriately reduces the associated variability. Use of multiple imputation (MI) is still rare in meta-analysis but if it is used, then the missingness is assumed to be missing at random. Further methodological research is needed to assess the functioning of MI especially given the weighting typically used in meta-analysis, however, it would seem likely to function best as a method for handling missing meta-analytic data.

Meta-analysis is also criticized for another form of missingness peculiar to this technique, namely, missingness due to *publication bias*. Publication bias is a term that refers to the scenario where only studies with statistically significant results are reported ("published") and only studies that are published (i.e., available) can provide data that can be synthesized in a meta-analysis. Clearly this kind of missingness will bias resulting ESs. There are many different ways researchers use to assess whether publication bias might exist. Graphical displays, such as the funnel plot are sometimes used. ES estimates based on smaller sample sizes would be expected to vary more than for studies based on larger sample sizes although the average ES should not depend on sample size. Funnel plots involve graphing ES estimates against their associated sample sizes and provide a graphical way of assessing whether this pattern holds. If the plots are skewed, then this might be inferred as evidence of publication bias although other explanations are also possible.

Indices are also available to assess potential publication bias. The fail-safe number as well as modifications thereof, trim-and-fill estimates can also be used to evaluate the potential for publication bias. Last, some meta-analysts use inferential tests of publication bias (e.g., Begg's rank correlation test, Egger's regression, and funnel plot regression). The reader is strongly encouraged to refer to any of the meta-analytic texts (especially Cooper & Hedges, 1994, and Rothstein et al., 2005) to find out further details about these different procedures. Meta-analysts are encouraged to use multiple methods for assessing publication bias including at least the trim-and-fill method as well as one of the regression methods despite their limited statistical power.

Meta-analysts should try to contact the primary study authors to obtain information that might not have been reported. In the absence of this information and if evidence supports the possibility of publication bias, meta-analysts are encouraged to use any of the variety of methods available for correcting for publication bias. In particular, the trim-and-fill correction and the use of weighted distribution theory-based approaches are strongly recommended.

11. Correction for Artifacts

Some meta-analysts use artifact correction procedures to correct for artifactual errors resulting from imperfect research scenarios. These correction procedures are designed to correct resulting ES estimates so that they represent results under ideal research scenarios (for example, they can be used to correct an ES estimate so that it represents the ES estimate based on perfectly reliable and valid test scores). The most commonly used correction is the correction for attenuation that can result from the lack of perfect reliability of scores on social science measures. Other corrections include correction for dichotomization of continuous variables and for restriction of range. Use of these procedures involves obtaining additional information (e.g., internal consistency reliability estimates for the relevant outcomes) to correct the relevant ES as well as its associated variance estimate. Use of artifact correction can also affect the sampling distributions assumed for the resulting ESs. Meta-analysts must specify which artifacts they might be correcting for and how. There is no consensus in the field about the use of these artifact correction procedures. Given the difficulties encountered in terms of gathering realistic values to calculate the corrections and their effects on the ESs' sampling distributions, the validity of the resulting corrections and of analyses conducted using the corrected ESs seems questionable.

12. Homogeneity of ESs

Meta-analysis is used to synthesize results from a multitude of studies designed to assess the same research question. While replication is encouraged in research, most studies do not exactly mimic each other. Studies tend to involve some subtle (or not so subtle) variation on a previous but similar study. Samples from different populations might be used (e.g., adults versus adolescents or college students, clinical versus non-clinical respondents, populations with different demographic information). Different implementations of an intervention might be tested. Different measures of a related but distinct construct might be investigated. This means that the resulting effect sizes might not come from a single population (sampling distribution of effect size estimates) with a single true effect size. Instead, it is more likely that while some of the variability in effect size estimates is due to sampling error, some of the variability is also attributable to random effects. In other words, the estimates do not come from a single population.

Meta-analysts should test the heterogeneity of the effect size estimates they gather. Methodological researchers have consistently supported use of the *Q*-test statistic designed to test the null hypothesis of homogeneous ESs. If the variability in the effect sizes is found to be more than could be solely attributed to sampling error, then this affects the model that should be assumed when conducting ensuing statistical analyses. Excess heterogeneity means that a random effects model should be assumed. If the effect sizes can be assumed homogeneous, then a fixed-effects model can be assumed. A meta-analytic researcher should clearly identify which model was assumed for all analyses including estimation of both pooled estimates as well as for analyses designed to investigate sources of variability in effect size estimates using the moderating variables detailed in Desideratum 5. Note that when reporting meta-regression model results, the choice made between a fixed- versus mixed-effects model should be clarified.

13. Descriptive Statistics

Meta-analysts should describe the resulting data that were gathered. This includes the availability of sample and study descriptors as well as information that could be used to calculate ESs. Some meta-analysts provide a table listing each study and associated descriptive information (such as the sample size underlying an ES as well as other study and sample descriptors as noted in Desideratum 5). This table usually also provides every ES or an overall ES for each study (see Desideratum 9). All meta-analysts present ES estimates pooled across studies for each outcome of interest and usually for levels of categorical moderating variables of interest. Along with all pooled estimates, associated standard error (and/or confidence interval) estimates should be provided. The (random-, mixed-, or fixed-effects) model that is assumed for these should already have been noted (see Desideratum 12).

14. Inferential Statistics

Results summarizing the tests of relation between moderators (see Desideratum 5) and the ES should be presented. Meta-analysts testing a number of moderating variables should consider use of (weighted) meta-regression model for testing the concurrent inter-relations. Conducting a multitude of statistical tests can lead to inflated Type I error for meta-analytic data as with any other kind of data. Controls such as the use of Bonferroni's correction to the nominal alpha level should be considered. Last, meta-analysts should appropriately model the meta-analytic data's structure. For example, in a meta-analysis involving multiple ES estimates per study, some of the moderators might be sample-level descriptors while others might be at the study level. The same considerations about centering of predictors (moderators) as those addressed with multilevel modeling of primary study data still apply.

15. Practical Significance

As with any empirical study, detection of statistical significance (or non-significance) should be interpreted within a context. While some researchers might cite rules of thumb for cutoffs representing small, moderate, and large effect sizes, interpretation of an effect size's magnitude should be made in the explicit context in which the effect size is calculated. For example, an ES estimate of 0.001 would qualify to most researchers as minuscule. However, in a test of aspirin for reducing heart attacks an ES estimate (R^2) of 0.011 was deemed sufficiently large that the trial was prematurely halted to stop "harming" placebo recipients who were not being given the aspirin (cited in Rosenthal, 1994). Thus, rules of thumb for describing an effect's size should be used with caution. Instead, the researcher should consider the magnitude and direction of the ES estimates in the context in which they are being assessed. Similarly, the strength (and direction) of the relation between the moderating variables and the ES should be interpreted at a practical rather than solely a statistical significance level.

References

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

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York: Guilford Press.

Cooper, H. M., & Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis*. NewYork: The Russell Sage Foundation.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *Handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

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

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

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

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

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

Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 177–203). New York: Russell Sage Foundation.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: John Wiley and Sons.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

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