Introduction to Meta-Analysis

Nazım Çoğaltay and Engin Karadağ

Abstract As a means to synthesize the results of multiple studies, the chronological development of the meta-analysis method was in parallel to a variety of definitions in the literature. Meta-analysis can be defined in different ways: as a means of summarizing and combining the quantitative results of research or as a method used to reach the quantitative effect size based on individual studies. Meta-analysis uses many quantitative approaches and calculation formulas when compiling multiple research findings. In this sense, no researcher needs to be an expert in all types and calculation formulas for all types of meta-analysis. However, if the researcher lacks familiarity with at least some of the main concepts of meta-analysis, then the correct results may not be obtained. This chapter aims to explain some of the main concepts of meta-analysis.

1 Introduction

The question of how to bring together and interpret research studies that are independent from one another is a basic and important question in all sciences. Hence, the inability to conduct research studies with large samples to represent a wider population because of obstacles such as time, cost and expert researchers and the discussion of how effective the findings of a single study can be have necessitated the synthesis of the results of a multitude of studies. The inadequacy of the results of a single study and the need to synthesize findings by scientists have led to the development of methodologies that allow for combining the results of many independent studies.

Many methods have been used to synthesize the findings of multiple studies. The first attempts at synthesizing studies can be observed in the efforts made to merge

N. Çoğaltay (✉)
Muş Alparslan University, Muş, Turkey
e-mail: n.cogaltay@alparslan.edu.tr

E. Karadağ
Eskişehir Osmangazi University, Eskişehir, Turkey
e-mail: enginkaradag@ogu.edu.tr

E. Karadağ (ed.), *Leadership and Organizational Outcomes*,
DOI 10.1007/978-3-319-14908-0_2

findings in the fields of astronomy and physics. Subsequently, experts in the field of agriculture began to develop statistical techniques that would allow for the compilation of repeated measurements (Hedges & Olkin, 1985). The compilation of data from multiple studies was conducted by means of narrative compilations. An expert in the field would read a study on a particular topic, summarize the findings and provide a conclusion regarding the summary of findings. However, this method was deemed ineffective because of limitations such as the varying subjectivity of different researchers (criteria, reliability, and validity) and the fact that only studies with a consistent effect size could be compared. These limitations of the narrative compilation method motivated scientists to seek a different methodology, and as a result, the methods of systematic review and meta-analysis emerged (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Systematic review and meta-analysis are two approaches aimed at synthesizing different studies that are independent of one another but also compatible. When both methods are used together, it is possible to compile the quantitative evidence, analysis and scientific approaches as a whole. This approach makes it possible to obtain a large sample size and to provide new perspectives on developing social policies. However, these two approaches are not synonymous; they represent two different approaches. Many meta-analysis studies are not systematic reviews. Meta-analysis studies can be a part of a systematic review, but this is not true of all meta-analyses (Littel, Corcoran, & Pillai, 2008).

It is believed that the first meta-analysis study was conducted by Karl Pearson in 1904 when he attempted to synthesize the independent vaccine studies concerning typhoid (Littel et al., 2008). However, it was not until the 1970s that social and behavioral scientists began using meta-analysis. Glass (1976) coined several statistical terms for synthesizing the results of more than one study. Studies from that period aimed to synthesize the results of independent studies on topics such as the effects of psychotherapy (Smith & Glass, 1977), the effects of classroom populations on achievement (Glass & Smith, 1978), the effect of interpersonal expectations (Rosenthal & Rubin, 1979) and the validity of race-based employment tests (Hunter, Schmidt, & Hunter, 1979). After the 1980s, scientists began to develop statistical methods or meta-analysis (Cooper, 1998; Cooper & Hedges, 1994; Hedges & Olkin, 1985; Light & Pillemer, 1984), and thus, meta-analysis became a statistical technique.

As a means to synthesize the results of multiple studies, the chronological development of the meta-analysis method was in parallel to a variety of definitions in the literature. Glass (1976), who first proposed the concept of meta-analysis, discussed primary analysis, secondary analysis and meta-analysis concepts and emphasized that these types of analyses were not to be confused with one another. He defined primary analysis as the analysis conducted in an original study, defined secondary analysis as the use of statistics to better understand the problem discussed in the original research or the use of data to find answers to new problems, and defined meta-analysis as the analysis of analyses. Meta-analysis can be defined

in different ways: as a means of summarizing and combining the quantitative results of research (Glass, McGaw, & Smith, 1981) or as a method used to reach the quantitative effect size based on individual studies (Durlak, 1995). The meta-analysis method differs from other quantitative review methods that attempt to test the correctness of hypotheses (Littel et al., 2008). Meta-analysis is the method of conducting a statistical analysis of the research findings of many independent studies conducted on a certain topic (Borenstein et al., 2009; Cohen, Manion, & Morrison, 2007; Glass, 1976; Hedges & Olkin, 1985; Littel et al., 2008; Petitti, 2000).

Meta-analysis uses many quantitative approaches and calculation formulas when compiling multiple research findings. In this sense, no researcher needs to be an expert in all types and calculation formulas for all types of meta-analysis. However, if the researcher lacks familiarity with at least some of the main concepts of meta-analysis, then the correct results may not be obtained. This chapter aims to explain some of the main concepts of meta-analysis.

2 Effect Size and Types

The main objective of the meta-analysis method is to determine a summary effect size by synthesizing data from multiple research studies. The effect size in meta-analysis is a measure of the strength and direction of the relationship between variables (Littel et al., 2008). This term may be expressed in different ways for various fields. In the field of medicine, the effect size is expressed as the application effect and is sometimes expressed as the odds ratio, the risk ratio or the risk difference. In social sciences, the term ‘effect size’ is used frequently but is sometimes expressed as the standardized mean difference or relationships (Borenstein et al., 2009).

The most frequently used effect size calculations fall into these categories: (1) proportions, (2) averages and (3) correlation coefficients. There is more than one way to calculate effect size in these categories. The preferred calculation of effect size will differ according to the aim and design of the study and the data format. Studies testing the effect of an intervention or studies aiming to make a variety of causal inferences (between pre- and post-test or between groups receiving and not receiving treatment) are in the category that use proportions and averages. Studies investigating the relationship between variables, besides causal direction inferences, are in the category of correlational meta-analysis (Littel et al., 2008). In other words, if the results of the effect size are numerical, then averages are used; if the results are nominal, then proportions are used; and if the results show a relationship, then correlations are preferred (Cohen et al., 2007). In addition,

it is also possible to classify meta-analysis studies into one of two categories: (1) comparison of groups and (2) correlational meta-analysis (Durlak, 1995).

There are two important differences in the calculations of effect size: dichotomous data and continuous data. Dichotomous variables are based on only two categories and frequently represent the presence or lack of a feature or situation. Pregnancy, high school graduation, and gender are examples of such variables. Continuous variables can have a range of values that can be expressed on a numeric scale. Examples of such variables include the number of pregnancies, the duration of training, and the duration of hospitalization. Test and scale results such as achievement tests or depression inventories can be considered continuous variables (Littel et al., 2008).

3 Effect Size in Dichotomous Data (Proportioning)

The effect size of dichotomous results is based on whether a phenomenon was observed. The most frequently used effect size measures are the *odds ratio (OR)*, the *risk ratio (RR)* and the *risk difference (RD)*. The odds ratio is the expression of the comparison of whether something has a probability of occurring (Littel et al., 2008). That is, the effect size is obtained from the proportion of two possibilities (Borenstein et al., 2009). The risk ratio, similar to the odds ratio, pertains to risk and is the ratio of risks to one another. The risk difference is the difference between two risks. The effect size of the odds ratio or the risk ratio is reached by converting data into logarithmic data, and the risk difference uses raw data to calculate the effect size. The odds ratio is the proportioning of the ratio of whether a certain phenomenon is observed in the experimental group to whether the phenomenon is observed in the control group. These effect size calculations are generally used in the fields of health and agriculture (*for more information, please see* Borenstein et al., 2009; Hedges & Olkin, 1985; Kulinskaya, Morgenthaler, & Staudte, 2008; Petitti, 2000). A hypothetical example showing calculations of the effect size of dichotomous data is shown in Table 1 (Littel et al., 2008).

Table 1 Effect size for dichotomous data in a hypothetical data table

	Event	No event	Total N	Odds	Risk
Experiment	4	6	10	4/6	4/10
Control	2	8	10	2/8	2/10
Odds ratio (OR) = (4/6)/(2/8) = 2.67 Risk ratio (RR) = (4/10)/(2/10) = 2.0 Risk difference (RD) = 0.40 − 0.20 = 0.20					

4 Average Effect Size Between Groups for Continuous Data

The effect size obtained from continuous data can be divided into two main categories: (1) the non-standardized mean difference (D) and (2) the standardized mean difference (d) or (g). Of these two types, raw data are used to calculate D means, and d or g is calculated using standardized techniques to convert raw data into other forms. These mean difference effect sizes are calculated using different techniques for each of the categories of data obtained from mean differences between groups independent of one another and from differences between the pre- and post-tests in the same group or matched groups (for further information concerning the techniques used, please see Borenstein et al., 2009; Hedges & Olkin, 1985).

The non-standardized mean difference (D) is used when all of the research included in the study is reported using the same scale. In such cases, meta-analysis is conducted by calculating the raw differences of the direct means to determine the effect size. However, the standardized mean difference (d) or (g) is used when results are reported based on different scales or methods in the studies included in the analysis. To compute the standardized mean difference, the resulting data are calculated by standardizing the standard deviation to equal 1 within the groups (Borenstein et al., 2009; Hedges & Olkin, 1985; Kulinskaya et al., 2008; Littel et al., 2008).

5 Correlational Effect Size for Continuous Data

The relational values obtained from research reporting the relationship between two continuous variables are the calculated effect sizes. The effect size of studies is generally obtained by calculating the Pearson correlation coefficient, *r*. Studies that provide this coefficient or that provide the opportunity to calculate this coefficient are included in the analysis. As this correlation coefficient is a value between +1 and −1, calculations are performed by transforming the *r* value into its corresponding *z* table value. The correlation coefficient is itself considered the coefficient of effect size and is also symbolized by *r* (Borenstein et al., 2009; Hedges & Olkin, 1985; Littel et al., 2008).

The effect width is considered when interpreting the effect size. This effect width is categorized in many different ways by various researchers; however, the most important categorization belongs to Cohen (1988), as shown in Table 2.

Table 2 Cohen's (1988) classification of effect width

Es metric	Small effect	Medium effect	Large effect
OR	1.5	2.5	4.3
SMD	0.2	0.5	0.8
r	0.1	0.25	0.4

OR odds ratio, *SMD* standardized mean difference, *r* correlation coefficient

6 Choice of Model

There are two main models used in meta-analysis studies: the (1) fixed effect model and the (2) random effect model. When deciding which model to use, the researcher must assess the characteristics of the research to determine which of the models' pre-conditions the study meets. In general, these two models use different processes to calculate the weights of studies, the average effect size and the confidence intervals for the average effects when calculating the effect size (ES). Therefore, to obtain the correct results in the processes of meta-analyses, it is important to choose the correct model in relation to the characteristics of the specific studies involved (Borenstein et al., 2009).

The fixed effect model has the (1) same assumption as the function of the research and (2) the aim of calculating only the effect size for the population. If it is determined that the function of the research is the same, that it shares a real effect and that the calculation of the real effect is not supposed to be generalized to wider populations, then the choice of model should be the fixed effect model. For example, a pharmaceutical company intended to conduct a drug trial study with 1,000 patients but has only been able to research one patient group at a time. Thus, the research was conducted more than once with repeated tests. In such cases, the model to be used to compile the repeated tests is the fixed effect model because the study was conducted by the same researchers and used the same doses and tests in patients from the same sample pool. Thus, all studies share the same real effect and meet all conditions for the fixed effect model, as the effect of the drug is investigated only in the identified population. It is important to note that it is uncommon to find meta-analysis studies of this type. It is nearly impossible to find research studies that meet the pre-conditions of the fixed effect model, especially in the social sciences and educational sciences.

In regard to the random effect model, it is assumed that the effect differs between sample groups and among studies. In summary, if the conditions of the fixed effect model are not met, then the random effect model should be used. The effects can differ in relation to the variables in the studies, such as the health, age, and education status of the sample subjects. For example, the effect size for a practice in the field of education may show variation among factors such as students, classroom populations and ages. In such cases, the appropriate model for meta-analysis is the random effect model.

It is important for a meta-analysis to correctly identify which model should be used for which type of research. As noted above, the choice of model should be made after identifying which pre-conditions are met by the studies. Borenstein et al. (2009) argued that to select a model based on the results of the heterogeneity test or to use the fixed effect model followed by the random effect model for the meta-analysis is not the correct approach and should be criticized. Further, the belief that the fixed effect model results in a stronger analysis is completely false. Therefore, it is not appropriate for researchers to use the fixed effect model under the assumption that it provides stronger results. The correct process is to select a

model by ascertaining which features of the studies included in the meta-analysis meet the pre-conditions of the model.

6.1 Heterogeneity

A heterogeneity analysis is the measure that shows how the effect width differs from study to study. This statistic tests whether the effects found by the different studies are caused by a sampling error or by a systematic difference between the studies in addition to a sampling error (Hedges & Olkin, 1985). The different effect sizes of the studies included in the meta-analyses make it necessary to find the size of the variance between the distributions. Therefore, heterogeneity tests are conducted to determine the conformity of the normal distribution of effect sizes. The impact value observed between studies show differences for two reasons. The first reason is the real heterogeneity of the effect size, and the second reason is related to errors within the studies. If researchers do not seek to test the heterogeneity, then they must separate the observed differences between the two components and focus on the first situation above (Borenstein et al., 2009).

The most common means of testing heterogeneity and determining whether the heterogeneity is statistically significant is the Q (df) statistic based on the χ^2 test. Structurally, all studies establish and test a null hypothesis to argue for a shared common effect (Hedges & Olkin, 1985). Under the null hypothesis, the Q value should follow the degrees of freedom equal to $k - 1$ and the central χ^2 distribution. When the effect sizes are heterogeneous, a statistically significant χ^2 value shows that the studies have different distributions and thus do not share a wide effect (Hedges & Olkin, 1985). The Q calculation formulas for meta-analysis studies are complimentary and homogeneous to one another and can be calculated in three different ways. Although all studies use Q_{Total} to test the common effect size (that is, the heterogeneity), $Q_{Between}$ is used to test heterogeneity between studies, and Q_{Within} is used while testing the heterogeneity within each particular study. There is an equality in $Q_T = Q_B + Q_W$ (Hedges & Olkin, 1985).

It is possible to test heterogeneity using several statistical techniques. The most common technique involves the Q statistic and is the sum of weighted squares, which aims to find the significance level of the differences observed in studies. T^2 is the variance of real effects. This value is used to calculate the weightings of studies under the random effect model. T is the standard deviation of real effects and is the same as the standard deviations of the effects of the same tests. This coefficient is used to predict the real effect distributions and is used when considering the important effects of these distributions. I^2 is the actual ratio of the observed distributions. The effects are not dependent on testing and can range in value from 0 % to 100 % (Borenstein et al., 2009).

7 Publication Bias

One of the components of greatest interest to researchers in meta-analysis studies is the effect of variance on the results observed. Have publication bias, the study design, sample characteristics or moderator variables influenced the observed effect? The identification of these or similar variables that have played a role in the resulting effect is important for meta-analysis and assists in the determination of correct results. This section attempts to explain the importance of publication bias in meta-analysis studies and how it is identified in meta-analysis studies.

Publication bias is based on the assumption that not all studies on a particular topics are published. Because studies that do not find statistically significant relationship or that find only a weak relationship are deemed unworthy of publication, they are believed to negatively affect the total effect or to create bias in increasing the average effect size (Borenstein et al., 2009; Kulinskaya et al., 2008). This publication bias effect, which can also be considered missing data, has a negative impact on the total effect of a meta-analysis. Therefore, publication bias should be considered in meta-analysis studies. To examine the publication bias of a study, researchers should consider the following questions (Borenstein et al., 2009):

- Is there any evidence of publication bias?
- Is it possible that the general effect size is the result of publication bias?
- To what degree is the total effect due to publication bias?

To answer the above questions using statistical methods, a series of calculations are used in the meta-analysis. One of the most popular of these methods is the funnel plot method. The figure obtained with this method may not be completely objective, but it provides the opportunity to determine whether publication bias affects such studies. A funnel plot conducted for a meta-analysis is shown in Fig. 1.

In the funnel plot above, there is no evidence of publication bias for the studies included in the meta-analysis. To speak of a publication bias, the funnel plot would need to present a serious degree of asymmetry. If a concentration of studies were plotted at the bottom end of the funnel below the line indicating the average effect size and skewed to one end (especially toward the right side), then a publication bias would be evident. The figure of a funnel plot can be interpreted as not representing serious publication bias for the effect size of the related studies.

Statistical techniques in regard to publication bias are not limited to the funnel plot technique. The more frequent use of the funnel plot may be explained by the practicality in its application and the visual aspect. In addition, one of the other techniques developed by Rosenthal (1979) is the *failsafe N* or the *file drawer number* technique. This technique assumes that it is possible to calculate the actual number of missing studies and argues that finding studies to include in a meta-analysis is necessary before determining whether the *p* value is significant. The use of this technique assumes that the main effect of missing studies have no effect. In addition, there is also the Duval and Tweedie *Trim-and-Fill* method (Duval &

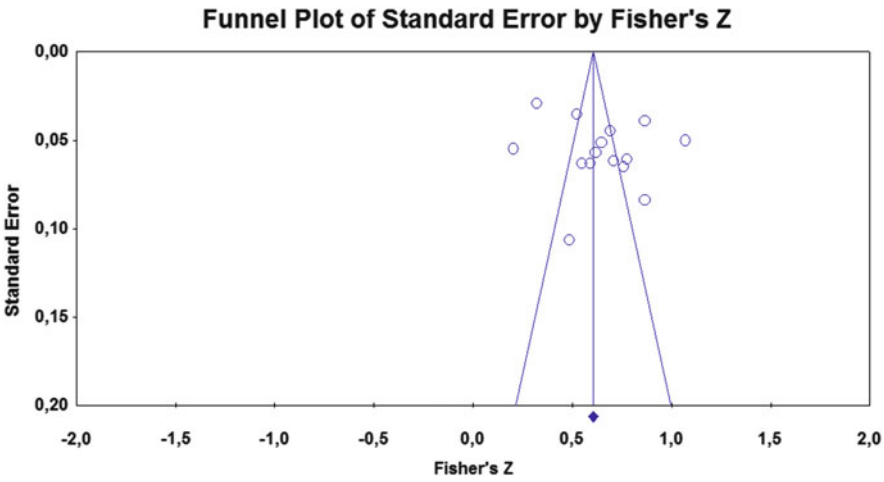

Fig. 1 Funnel plot of standard error by correlation coefficient (*r*)

Tweedie, 2000), which uses a repeated technique to remove small studies at the extreme ends of the positive end of the funnel diagram. The trimming and filling process is repeated until the funnel diagram is symmetric in regard to the effect size (Duval, 2005).

8 Sub-group Analysis and Moderator Analysis

A meta-analysis not only predicts the average effect based on all studies included in the analysis but also allows for the calculation of the average effects of various subgroups of studies and enables comparisons between these effects. Subgroup and moderator analyses are methods developed to test the statistical significance of differences between groups.

A subgroup analysis is a comparison of the effects of two or more groups. Three methods are used for the analysis of subgroups. A *Z* test is used to compare the average effect sizes of two groups, and a variance analysis or *Q* test is used to compare two or more groups. All three methods are based on mathematical formulas (Borenstein et al., 2009). Moderator analysis is an analysis method that attempts to test the differences between the average effect sizes of variables (moderators) and the direction of these differences. In a meta-analysis study, subgroup and moderator analysis are well planned in regard to the objective of the study, and the processes are conducted as planned (Littel et al., 2008).

The statistical significance between the difference of the subgroup analysis and moderator variables is tested using the Q statistic. In this method, Q is divided into two, as Q_{Within} (Q_w) and $Q_{between}$ (Q_b), and the analysis aims to find meaning based on the two Q values. Q_w attempts to test the homogeneity within the group or

moderator and determines whether the variance within the groups is statistically significant, Q_b tests the homogeneity among groups or variables and attempts to determine whether the variance between the groups is statistically significant, and Q_T determines whether the groups are statistically significant (Borenstein et al., 2009; Hedges & Olkin, 1985; Kulinskaya et al., 2008).

References

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillside, NJ: Lawrence Erlbaum.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. London: Routledge.

Cooper, H. (1998). *Synthesizing research*. Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L. V. (1994). *Handbook of research synthesis*. New York: Russell Sage Foundation.

Durlak, J. A. (1995). *Reading and understanding multivariate statistics*. Washington, DC: American Psychological Association.

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Bornstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 11–33). Chichester: Wiley.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463.

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

Glass, G., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, LA: Sage.

Glass, G. V., & Smith, M. K. (1978). Meta-analysis of research on the relationship of class size and achievement. *Educational Evaluation and Policy Analysis, 1*, 2–16.

Hedges, L. V., & Olkin, I. (1985). *Statistical method for meta-analysis*. San Diego, CA: Academic Press.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86*, 721–735.

Kulinskaya, E., Morgenthaler, S., & Staudte, R. G. (2008). *Meta analysis: A guide to calibrating and combining statistical evidence*. London: Wiley.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Littel, H. J., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York: Oxford University Press.

Petitti, D. B. (2000). *Meta analysis, decision analysis and cost effectiveness analysis: Methods for quantitative synthesis in medicine*. New York: Oxford University Press.

Rosenthal, R. (1979). The file drawer problem and tolerance of null results. *Psychological Bulletin, 86*, 638–641.

Rosenthal, R., & Rubin, D. B. (1979). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3*, 377–386.

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