This is chapter 2: Basic steps of meta-analysis and the emergence of approaches (pp. 9-15) from

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Hogrefe & Huber.

2

Basic Steps of Meta-Analysis and the Emergence of Approaches

Up to this point it has not been clearly stated how the stronger procedural and statistical rigor of meta-analysis in comparison to traditional reviews manifests itself. In this chapter, the basic steps of meta-analysis will be outlined. Meta-analysis is conceived as a process comprising several steps of which one — methods of statistically aggregating study results — is the main focus of this book. Before the statistical details will be presented in the next chapter, meta-analysis will be presented from a bird's eye view. The emergence of meta-analytical approaches is outlined subsequently.

2.1 BASIC STEPS OF META-ANALYSIS

It is useful to commence with the introduction of terminology. Most researchers are familiar with methods to analyze original data from an individual study. Such analyses will be called *primary analyses* in the present context. Another form would be *secondary analysis* which designates a reanalysis of existing data to apply different and supposedly better analytical methods and/or to answer new research questions (Glass, 1976). This latter form of data analysis will not be of concern in what follows.

Normally, the data in primary analyses results from measurements of person characteristics (individual units), like abilities, attitudes, and the like. A primary analysis is mostly conducted to describe these characteristics and/or relate them to or explain them by other variables. Thus, in a study on the predictive validity of an intelligence test for job performance, for example, a number of n persons participates in a study and provides a number of n pairs of observations for the two variables.[1] As a result of a primary analysis, the typical outcomes are a correlation coefficient for the two variables and a test statistic to make inferences to a population. In this case, the correlation coefficient is a measure of effect size, because it expresses the strength of the (linear) relationship.

Now consider that after publication of the results of the first study a second one on the same relationship is conducted. In the second study a new sample is drawn with a different number of n individuals, and the correlation is again computed in a primary analysis. Additionally suppose that the effect size in the second study is different from the first one. The question — typical for all literature reviews — arises what a good summary of both studies' results is in the given case. Further assume that the second study could be considered to be a replication of the first one. That is, the same measures were used, the sample was drawn from the same population, and so forth. Under these circumstances it would be reasonable to pool the data of both studies, if available, to arrive at a single effect size based on the total sample of both studies. Unfortunately, this is rarely the case and the task then still is to somehow summarize the effect sizes.

Taking this idea of additional study results on the same research question further, a situation is given that calls for an integrative review of empirical studies. Such a situation is illustrated in the lower and middle part of Figure 2.1 (Level 0). Here, different individual units are sampled in a number of k different studies on a common research question.[2] At this zero level, primary analyses result in empirical reports to be summarized, which include a number of (at least) k effect sizes. Figure 2.1 provides an illustration with correlations $(r_1, \ldots, r_k)$ as effect sizes — the main focus of the present book. Of course, it is not necessary to always collect pairs of observations at Level 0, nor is this process only applicable to correlations as effect sizes.

In a broad sense, meta-analysis is a systematic process of quantitatively combining empirical reports to arrive at a summary and an evaluation of research findings. This "analysis of analysis", as Glass (1976, p. 3) has defined it, can be located in the upper part of Figure 2.1 (Level 1). In analogy to primary analysis, it includes the statistical aggregation of individual units. In contrast to primary analysis, however, the individual units are aggregate measures resulting from Level 0 analyses. The result of a meta-analysis is symbolized only by θ in Figure 2.1. Much more will be said about such a pooled estimate of an effect size in the following chapters. Here, it suffices to say that one of the aims of most meta-analyses is to arrive at such a single summary measure.

Nevertheless, meta-analysis is characterized by many more attributes than simply a step of statistical aggregation. One of the other important attributes of meta-analysis is the more general call for a stronger procedural system-

[1]That is, a total of n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ are observed, where x_1 denotes the intelligence score of Person 1 and y_1 his/her job performance score in the example.

[2]Of course, sample sizes need not be the same in such studies, a fact that is not necessarily clear when inspecting Figure 2.1.

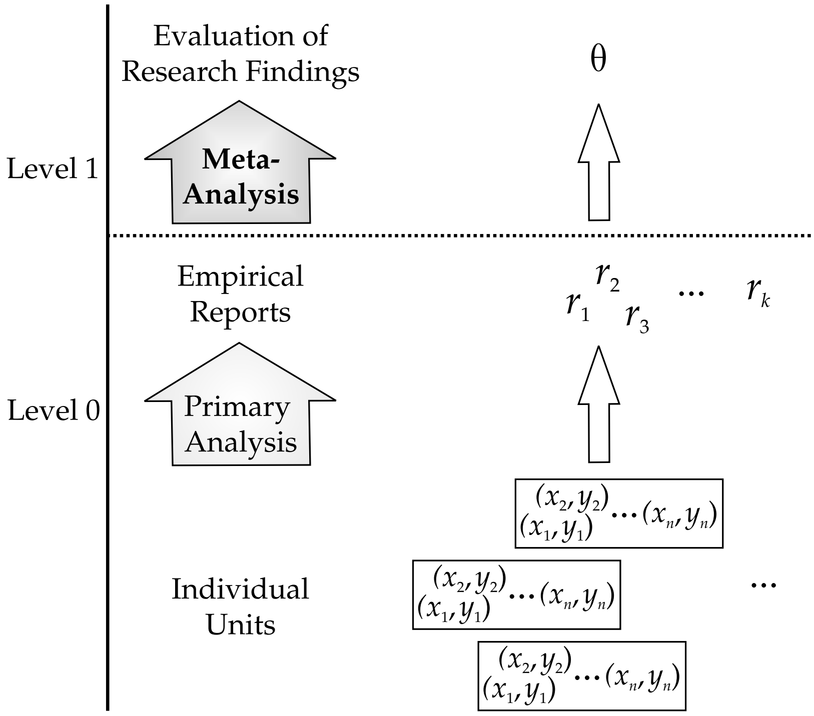

Figure 2.1 Different levels of analysis: Primary analysis and meta-analysis.

atic reviewing of the literature. Correspondingly, several guidelines for meta-analysis have been published. Some focus more on the whole process (e.g., Jackson, 1980), others rather give methodological guidelines (e.g., Cook, Sackett, & Spitzer, 1995). One widely accepted specification of the stages or conduct of a meta-analysis was presented by Cooper (1982; Cooper & Hedges, 1994a), which is formulated in close analogy to the stages of primary analysis:

1. Problem formulation
2. Data collection
3. Data evaluation
4. Analysis and interpretation
5. Public presentation

For each of these stages, attempts were made to clarify the questions to be answered and the methods to arrive at the respective solutions to problems posed. At every stage there is a demand of the meta-analyst for a maximum of explicitness. The whole process of reviewing has to be structured, and it has to be made reconstructible to the research consumer as to how the reviewer arrived at his conclusions. Thus, one of the main criticisms of traditional reviews is addressed by this requirement.

The first stage not only includes the tasks of clearly specifying the research question to be answered by a review and laying the foundations of exclusion and inclusion criteria for the studies to be synthesized, but also covers questions about what statistical model is to be assumed in a meta-analysis. The problems and corresponding solutions for the formulation of the research question to be answered are presented by Hall, Tickle-Degnen, Rosenthal, and Mosteller (1994) in detail and need not be repeated here. What is of much greater concern for the present study are the statistical models available for research synthesis. The available models will be presented and discussed in considerable detail in Chapters 4 and 5. It is important to note that first, statistical matters are not only questions about the proper formulae to use, they are also *conceptual* questions that cannot be answered on the sole basis of empirical results (Hedges, 1994b). Much of the theoretical as well as empirical parts of the present book are devoted to the explication of models and evaluating the performance of statistical procedures associated with different models when their assumptions are met or violated.

The second and third stages of the process concern data retrieval and its evaluation. Tasks and potential problems arising in connection with the former step are presented by White (1994) as well as Reed and Baxter (1994). Data evaluation is the task to judge the quality of the retrieved literature and (optional) assignment of quality scores to the studies under review that can be used in subsequent steps to weight the studies in the process of aggregation (Wortman, 1994). Of course, catalogues of quality criteria are essential at this stage and are available for research in medicine (Chalmers et al., 1981) and psychology (Shadish, Cook, & Campbell, 2002), for example. The idea guiding these two steps is to disclose procedures and criteria for data collection, selection, and weighting in the process of synthesizing. Traditional reviews are often criticized for not being explicit enough at these stages in particular.

Before collected data is actually aggregated, it has to be extracted from the available empirical reports. What is meant by the extraction of data will be detailed in Chapter 3 on effect sizes. The task to be dealt with here is to quantify the results of interest in a measure of effect size common to all studies under investigation. The quantification to be carried out aims at making the results of the studies amenable to statistical aggregation. This represents an essential part of a meta-analysis on the one hand, and another important difference to the narrative review on the other. Hence, meta-analyses are in general also more precise in results as compared to traditional reviews and enable the meta-analyst to make statements about the size of an aggregate effect and its significance. This goes beyond more vague summary statements ordinarily made in narrative reviews. For the last stage of presenting results of a meta-analysis there are also rather precise guidelines. Special forms of reporting meta-analytical results have also been developed. More information on this topic can be found in the works of Halvorsen (1994) as well as Light, Singer, and Willet (1994).

In sum, all these stages of meta-analysis can be characterized as an effort to more precisely structure the whole process of reviewing the literature, explicitly state the goals, and give guidance as to how to tackle with potential problems of each stage. Comparisons with traditional reviews, for example by Cook and Leviton (1980), are therefore strongly in favor of meta-analysis as the method of choice. An empirical comparison between meta-analysis and traditional reviews has been conducted by Beaman (1991) who also concludes that meta-analysis seems to be the preferable method.

What is quite clear from the preceding account is that meta-analysis is not yet another arcane set of statistical formulae a scientist has to deal with but a method to successfully treat the whole complicated process of synthesizing the scientific literature. There surely are a lot of steps in this endeavor that can be classified as qualitative rather than quantitative and these very aspects have mainly been the focus of critics of the method (e.g., Eysenck, 1978). Although the statistical methods of meta-analysis have also been the subject of several controversies (e.g., see Chapter 4) they were not the main target of fundamental critics.

In comparing primary analysis and meta-analysis, several similarities can be noticed. Of course, this is due to the process of meta-analysis being specified in analogy to primary analysis, as outlined above. This makes it quite easy to understand what meta-analysis actually is about and what its basic aims are. Taking a closer look at statistical aspects (Stage 1 and 4), things get more complicated because a higher level of abstraction from the original data is introduced. The statistical foundations of meta-analysis have been presented in various articles in a more concise form (e.g., Hedges, 1983a), introductory books (e.g., Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Lipsey & Wilson, 2001; Rosenthal, 1991) as well as handbooks (Cooper & Hedges, 1994b). What makes an acquisition of the techniques somewhat difficult for the uninitiated is the unfamiliar statistical data to deal with. Ordinarily, a researcher in the behavioral sciences applies data-analytical techniques to the results of an experiment or observational study (Level 0 in Figure 2.1). A number of individual units provide measurements on a set of variables of interest, with measurement instruments chosen to represent the true scores of the persons on the variables as reliably and validly as possible. If a researcher aims at testing certain theoretical propositions, the size of prespecified relationships between the (observed) set of variables is estimated and tested by using data resulting from the measurement process. Estimation and tests in this context are conducted to arrive at statistically well-founded propositions about the relationships of interest in a *population* of *persons*. These outcomes, the estimate and test results, constitute the data basis of meta-analysis. The meta-analyst therefore does not directly deal with measurement of persons but results from studies which can be viewed as aggregated measurements. As a result, the objects of examination are studies and not persons, and the inference the meta-analyst aims at is not from a group of persons to a population of persons but from a group of studies to a *universe* of studies. Analogously to the situation in primary analysis, the empirical reports collected at Stage 2 are conceptualized as a sample of studies from a larger universe of studies. Inference in meta-analysis refers to such a universe, and one of the most difficult questions to be answered in meta-analysis is how this universe can be conceptualized or characterized — though it might be noted that a specification of the population in primary analysis is not an easy to answer question either (Frick, 1998). Variants of universe characterizations will be presented in Chapter 4 in detail, so their discussion is postponed till then.

To summarize, the principles of applying statistical techniques almost remain the same for meta-analysis, but an additional level of abstraction is introduced. The transfer of questions arising in the context of primary analysis to meta-analysis is helpful for understanding the method and raising critical questions in its application.

2.2 ON THE EMERGENCE OF APPROACHES

At least in the field of psychology there are some obvious peculiarities stemming from the history of meta-analysis in this field (for an interesting and comprehensive overview, see Hunt, 1997). In the early 1980s several proponents of meta-analysis presented comprehensive treatments of the subject (e.g. Glass et al., 1981; Hunter, Schmidt, & Jackson, 1982). As mentioned in Chapter 1, these collateral specifications of meta-analysis were concerned with developing methodological solutions for vastly different substantive problems. On the one hand, Glass and colleagues dealt with (quasi-)experimental designs on the comparison of psychotherapies, and Hunter and Schmidt were concerned with the problem of predictive validity in personnel selection. Hence, the former focused on methods to aggregate mean differences and the latter on correlations. Furthermore, there have been specific features in these areas of application that have caused different accentuations. For example, in the area of personnel selection it is customary to apply corrections to the correlation coefficient for range restriction in the sample. This is due to the fact that at least for one of the two variables to be correlated (job performance, for example) only scores of a subsample of the total applicant pool are available. Hence, the treatment by Schmidt and Hunter considers such corrections as being of utmost importance, and a large part of their methodological contributions to meta-analysis is concerned with them, whereas those of others are not.

These two groups of authors are not the only ones who have presented comprehensive treatments of meta-analysis in the psychological literature. Again, additional presentations have a somewhat different focus. Rosenthal (1978) presented methods for the combination of probabilities as study results and was the first to consider the so-called *file-drawer problem* in meta-analysis in depth (Rosenthal, 1979). The file-drawer problem refers to the suspicion that in the behavioral sciences the publication of significant results is favored by editorial policies and journal reviewers' evaluations, thereby causing a biased sample of study results to be available to the meta-analyst. Another major effort — if not the most detailed and statistically elaborate in the behavioral sciences to date — to specify the (statistical) methods of meta-analysis was presented by Hedges and Olkin (for a comprehensive overview, see Hedges & Olkin, 1985). Here, the main focus was not a substantive problem, but a precise statistical formulation of the models in meta-analysis and the presentation of corresponding proofs for the situations given in meta-analyses.

In sum, different groups of authors with different substantive and technical focus have dealt with the methods of meta-analysis — many of them simultaneously — to arrive at a pre-packaged comprehensive treatment of the topic. Such packages, associated with different author names, focus, and procedures, will be called approaches in the following. The publications corresponding to the approaches soon became standard references in certain subdisciplines in psychology. For example, the work of Hunter et al. (1982) became a quasi-standard in the field of industrial and organizational (I/O) psychology, whereas the work of Glass et al. (1981) was the main reference for meta-analytic research in educational psychology. It quickly became accustomed to researchers from different areas to rely on these different approaches. They also became deeply entrenched in research habits in certain subdisciplines. Thus, many researchers either thought that the application of the approach most pertinent in their field of study was the only (correct) option (e.g., Huffcutt, 2002), or the choice of an approach would be inconsequential for the results, or even that differences between the approaches in recommendations, treated effect sizes, and formulae were perhaps simply another mystery of statistical methods in the social sciences.

Several different approaches are identified in the psychological literature. As might be suspected, classifications of proposed techniques into approaches do not always fully agree. For example, a trio of meta-analytical approaches is identified by Andersson (1999) as well as Johnson, Mullen, and Salas (1995), but other categorizations have also been made (Bangert-Drowns, 1986). What is important in the present context is that different approaches still coexist in the psychological literature and their differences are at least partly due to historical reasons, specifics of the substantive research question, and only rarely on diverging mathematical-statistical derivations.

It is interesting to note that in a field like medicine, where scientists adopted the methods of meta-analysis considerably later and with more reservations as compared to psychology, such differences in approaches hardly exist. When inspecting overviews in medical research (e.g., Sutton et al., 2000), nothing comparable to the situation in psychology can be recognized, and the focus is more on statistical models rather than substantive questions.

To summarize, developments of the methods of meta-analysis are different in diverse fields, they were influenced by historical and substantive aspects, and specific approaches are almost tied to different subdisciplines in psychology. Finally, it should be added however, that the differences between approaches concentrate on their procedural recommendations. That is, study retrieval methods (Stage 1), data evaluation (Stage 3), and public presentation format (Stage 5) are highly similar. The differences can be located at Stage 1 in the formulation of the statistical model and Stage 4, the analysis procedures. The following presentation therefore focuses on these aspects in comparing the approaches. A detailed theoretical comparison is given in Chapter 5, and the comparative quality of results is assessed in a Monte-Carlo study to be presented in Chapters 7 and 8.