

Chapter 17

Focus Chapter

META-ANALYTIC RESEARCH METHODS

JOSEPH A. DURLAK, PH.D.

This chapter cannot tell you all you need to know about how to do a meta-analysis; it can, however, inform you about *when* to do one. This slightly different orientation permits discussion of several critical issues useful for those who initiate their own meta-analysis as well as those who want more information about how to evaluate meta-analyses that appear in the literature. If your knowledge of meta-analysis is rudimentary, see Light and Pillemer's (1984) excellent nontechnical introduction to meta-analysis and Rosenthal's (1995) highly readable piece on how to write a meta-analytic report. Explanations of the most common technical and procedural aspects of meta-analysis with plenty of helpful examples are available in several other sources (Cooper & Hedges, 1994; Durlak, 1995; Durlak & Lipsey, 1991; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Wolf, 1986). The first and last three of these sources are texts devoted to meta-analysis.

Before we proceed, note that some researchers would answer the question "When should one do a meta-analysis?" with an emphatic "Never!" Although an estimated 1,500 meta-analyses have appeared in print and approximately 100 new meta-analyses appear each year in the social sciences, controversy still swirls around the technique.

Sharpe (1997) offers a good commentary on the pro and con arguments about meta-analysis. Well done reviews are extremely helpful in synthesizing past research and highlighting potentially fruitful directions for new work. Poorly done reviews, however, can be harmful if they offer misleading information and unjustified conclusions that might close off inquiry in some aspects of a field prematurely or misdirect researchers to the wrong issues and variables. It is true that meta-analysis can be misused and misinterpreted, but so can any research method and statistical technique. This chapter is written from the perspective that, in principle, meta-analysis is an appropriate research strategy, although in practice, meta-analyses definitely vary in their quality, utility, and overall value.

Perhaps the most common misperception of meta-analysis, which is generated by its statistical features (after all, numbers don't lie, do they?), is that meta-analysis is a highly structured objective strategy with clear decision rules for each important step. Actually, there is no one standardized approach in meta-analysis. Several advances in meta-analytic techniques have occurred since Smith and Glass (1977) popularized the use of meta-analysis within the social sciences with their review of psychotherapy research. Moreover, many decisions are required while conducting a meta-analysis. It is extremely important that these decision points be made explicit so that the quality of a

meta-analysis can be judged. Some of these judgment calls are discussed here, and further information is available in several sources (Matt, 1989; Nurius & Yeaton, 1987; Wanous, Sullivan, & Malinak, 1989).

WHAT IS META-ANALYSIS?

Meta-analysis is an approach to research synthesis whereby the results of different studies are transformed into a common metric, the effect size, which is then pooled or aggregated across studies. Meta-analysts usually report the overall mean effect obtained from all reviewed studies and for important subcategories of studies and then attempt to explain these outcomes by searching for moderators. Studies always vary in their outcomes; if every investigation obtained similar results, there would be little need for a review in the first place. In other words, a major question in most meta-analyses is: What accounts for the variability of obtained effects?

There are two main types of effect sizes (ESs): product-moment correlations and standardized mean effects. Because meta-analyses of treatment studies usually use ESs, this chapter only discusses these types of effects, which are also referred to as d , d_t , or g . However, product-moment correlations can be used effectively for reviewing clinical research, as demonstrated in Reid and Crisafulli's (1990) meta-analysis. Across 33 studies, there was an average r of 0.16 between the extent of marital discord present in the home and the level of boys' externalizing problems. An r of 0.16 corresponds to a mean ES of 0.32.

In most treatment meta-analyses, an ES is calculated within each study by subtracting the posttreatment mean of the control group from the posttreatment mean of the treatment group and then dividing by the pooled standard deviation (Cooper & Hedges, 1994). The ESs from each study are then averaged to produce mean ESs.

How to Judge Mean Effects

Table 17.1 presents the distribution of mean ESs obtained in a meta-analysis of 156 meta-analyses of behavioral, psychological, and educational treatments (Lipsey & Wilson, 1993). These data indicate the magnitude of ESs typically obtained in the social

Table 17.1. Effect sizes obtained from 156 meta-analyses of behavioral, psychological, and educational treatments

Proportion of Meta-Analyses with Certain Effects	Magnitude of Effect
Mean of all meta-analyses	0.48
68% of all meta-analyses	0.19 to 0.75
16% of meta-analyses	> 0.75
16% of meta-analyses	< 0.19
5% of meta-analyses	> 1.00
0.006% of meta-analyses	< 0.00

Note: Data are drawn from Lipsey and Wilson (1993), Figure 7, and are based on treatment versus control group designs.

sciences and education. The mean ES drawn from all 156 meta-analyses was 0.48: in one-sixth of the reviews, the mean ES was greater than 0.75; in another one-sixth, the mean was less than 0.19; and only one meta-analysis reported a negative mean effect, which indicated that the control group did better than the treatment group over time. Lipsey and Wilson's (1993) data are consistent with Cohen's (1977) initial suggestions, offered before meta-analysis became popular, that mean ESs of 0.20, 0.50, and 0.80 should be considered small, moderate, and large in magnitude, respectively.

However, the magnitude of an ES does not necessarily reflect its practical significance. Much depends on what the outcomes are. In some cases, "small" ESs can have substantial practical value if they are based on such outcomes as the presence or absence of a clinical diagnosis, graduation from school, arrest rates, or serious antisocial behavior (Lösel, 1995). If the outcome is a matter of life and death, as it sometimes is in medical trials, ESs as low as 0.07 can nevertheless represent a highly successful intervention (Rosenthal, 1991). There are now several ways to calculate the clinical or practical significance of ESs (see Baucom & Hoffman, 1986; Durlak, Fuhrman, & Lampman, 1991).

ESs can also be calculated from single-subject designs (Busk & Serlin, 1992; see also Gaynor, Baird, & Nelson-Gray, this volume), from one-group-only designs (Lipsey & Wilson, 1993), and in studies comparing two or more treatments but lacking a control group. However, ESs should not be combined across these categories. Compared to ESs obtained from treatment versus control situations, ESs from single-subject designs are based on within- rather than between-group data, one-group-only designs are often much higher in magnitude, and treatment versus treatment comparisons are much lower in magnitude. In the second case, there is no control group that can change positively over time, and in the third case, another treatment should have more positive impact than a no-treatment condition.

Major Steps in a Meta-Analysis

There are six major steps in a meta-analysis, which are listed in Table 17.2. The steps include (1) formulating the research question(s), (2) doing a literature search, (3) coding studies, (4) making decisions regarding the calculation of effects, (5) conducting statistical analyses, and, finally, (6) offering conclusions and interpretations. Each step has several major parts that are described in detail elsewhere (Durlak & Lipsey, 1991). Step 4, for instance, involves such issues as dealing with multiple treatment

Table 17.2. Major steps in a meta-analysis

1. Formulate the research question(s)
2. Do an adequate literature search
3. Code relevant studies
4. Make decisions on calculating effects
5. Conduct statistical analyses
6. Offer conclusions and interpretations

Note: See Durlak and Lipsey (1991) for extended discussion of each step.

groups and outcome measures in individual studies, appropriate weighting procedures, and adjusting ESs based on small sample sizes.

This chapter focuses on five issues to consider in deciding when to do a meta-analysis that relate to four of the six steps. These issues are offered as questions to ask when planning a meta-analysis: Do you have specific hypotheses (Step 1)? Are there enough studies to code? How will you obtain representative studies? (Both are related to Step 2.) How will you code studies (Step 3)? How will you rule out rival explanations (Step 5)?

Do You Have Specific Hypotheses?

You should do a meta-analysis when you have specific hypotheses about a research literature that can be tested. Just as in an individual experiment, it is better to start with specific hypotheses concerning what you hope to find rather than embark on a fishing expedition that analyzes every possible variable and relationship. In the latter case, as long as you keep doing analyses, something is bound to come out significant. Unfortunately, there are examples of meta-analyses in which multiple analyses were conducted without any hypotheses to guide them (and without any control of Type I error), and in which authors have attached undue importance to a few significant results that appeared among the many analyses that were attempted.

An important requirement in doing a useful meta-analysis, which many do not realize, is the need for adequate working knowledge of the relevant research, including its major theories and procedures *prior to doing the review*. This step is important to formulate the best a priori hypotheses. For example, what controversies exist in the field? Can you identify certain conceptual, theoretical, or procedural issues that might account for disparities in research findings? Have new theories been developed or introduced that might be applied to past studies or research from another area? Are there new findings from well done investigations that suggest which variables might be most important? These are the types of issues that generate interesting hypotheses.

Are There Enough Studies to Review?

Suppose, in an individual study, we wanted to analyze the differences between a treatment and control group on an outcome measure using a *t*-test. The power of the statistical analysis in this case is determined by a combination of three factors: the effect size (which often has to be estimated), the probability level of the statistic being used, and the number of subjects in the two groups. If the population effect size were 0.50, the chosen probability level were 0.05, and a two-tailed test were conducted, we would need 64 subjects *per group* to have 80% statistical power, which in most cases would be sufficient.

In treatment effectiveness meta-analyses, mean ESs generated by group studies are often compared. For instance, the mean effects produced by one type of treatment are compared to the mean produced by another type of treatment. In this case, power is determined by the same first two factors (the effect size and probability level) and by the number of *studies* that are being compared, not by the number of subjects in these studies. If we used a *t*-test to compare group ESs, we would need 64 *studies* of each type of treatment to reach 80% power. Meta-analyses rarely have this degree of power throughout all their analyses.

Sufficient statistical power has been an important limitation in many meta-analyses and may be one explanation for the sometimes surprising finding that the type of treatment makes no difference in outcome, that clients with all types of problems improve equally with intervention, and so on. It is not unusual to have some analyses conducted on fewer than 10 studies per group. In our above example, 10 studies translates into less than 20% power. We cannot reach very strong conclusions under such limiting conditions.

How many studies are needed for a meta-analysis? This question can only be answered in reference to the specific aims of the meta-analysis. Although 128 studies, for example, would seem sufficient, it may not be once studies are subdivided to examine specific research questions. For instance, suppose one predicts that behavioral treatment will produce significantly higher ESs than dynamic treatment when specific measures of problem behavior are collected, but dynamic treatment will yield significantly better outcomes on measures of self-esteem and personality functioning *at follow-up but not at posttreatment*. The latter prediction might be made based on the logical premise that gains from dynamic treatment consolidate slowly and it is only some time after treatment ends that the true effects of treatment will be realized.

The above hypothesis would be impossible to assess in the child therapy literature. One can probably find 50 studies of behavioral treatment measuring outcomes on specific problems and on self-esteem at posttreatment, but there are not 50 behavioral studies with follow-up data on both these measures, and there are only a handful of dynamically oriented child therapy studies at all, much less ones with follow-up data. In other words, there is no point in attempting to do the impossible in a meta-analysis. The necessary data are sometimes unavailable.

Typically, the only way to discern if a meta-analysis is possible, or which hypotheses can be tested, is to do a lengthy and time-consuming literature search and then code the studies (see below) to determine if there are sufficient data for analysis.

How Will You Obtain Representative Studies?

Although power considerations suggest that one should collect as many studies as possible, there are two possible complications. First, decisions have to be made about what studies to include and exclude. Second, reliable methods must be used to find relevant reports. There is no such thing as an exhaustive search of the literature; there are simply too many studies in too many published and unpublished sources to identify every study done on any topic. The major issue is *representativeness*. How will you obtain a nonbiased representative sample of studies for review?

Identifying Relevant Studies The domain of eligible research must be described by explicit inclusionary and exclusionary criteria that operationalize exactly what research is being reviewed, the minimal requirements for sample inclusion, and the basis for excluding studies. For instance, in an attempt to describe several characteristics of child therapy outcome research, we searched for all studies "in which some form of psychotherapy for maladapting children (ages ≤ 13) was compared with a control group" (Durlak, Wells, Cotten, & Johnson, 1995). We went on to define what was meant by psychotherapy and excluded drug treatments, peer counseling, and family therapy.

The inclusionary and exclusionary criteria are guided by the meta-analyst's specific research questions. If one were interested only in the effects of treatment on children with certain types of problems, problems that reach a specific level of clinical severity, or clients treated only by professional therapists, then the search criteria would reflect these specific aims.

Finding Relevant Studies Do not depend solely on computer searches to secure an adequate sample of studies because such searches are notoriously unreliable in terms of their true positive hit rate. Computer searches tend to identify high numbers of irrelevant studies and frequently miss relevant reports. This is because the indexing system for computer searches rarely corresponds precisely to a reviewer's interests, no database is comprehensive, and some subjectivity is involved when each individual study is indexed.

Therefore, computer searches are not as simple as using a few terms to capture all relevant studies. In some cases, computer searches have captured less than 6% of the relevant literature (Lösel, 1991), or identified more than 10 times as many studies than eventually qualify for a review (Weisz, Weiss, Alicke, & Klotz, 1987). The Cooper and Hedges (1994) volume offers several helpful suggestions about literature searches.

Three search strategies are combined in many well done meta-analyses and these include computer searches, manual searches of journals that typically publish articles in the research area, and inspection of the reference lists of included studies and previous research reviews. This three-pronged approach is much more likely to yield a representative group of studies.

Publication Bias Meta-analysts must deal with the issue of publication bias, which refers to the reluctance of editors and reviewers to accept for publication studies with nonsignificant results coupled with, and this is frequently overlooked, the hesitation on the part of authors to submit their nonsignificant findings for possible publication. Basically, this means that published studies are more likely to produce higher ESs than unpublished reports, a phenomenon that is quite common in the social and medical sciences (Dickersin, 1997).

Unfortunately, unpublished studies are particularly difficult to track down because they may consist of dissertations, convention papers, technical reports, and studies lurking in investigators' file drawers. Researchers vary in their willingness to send in copies of their unpublished work, so there can be many "fugitive" studies that are never obtained for scrutiny.

Although they do not represent all of the unpublished literature, selecting dissertations for analysis is an appropriate strategy. Databases targeting dissertations and volumes of *Dissertations Abstracts* can be searched to estimate the population of relevant studies, and dissertations can be obtained from most institutions through interlibrary loan. In addition, dissertations often contain more procedural details than published studies.

The importance of clear inclusionary and exclusionary criteria and a careful search for representative studies is underscored when one compares different reviews of ostensibly the same literature. For instance, no single study appeared in each of four reviews of school-based drug prevention (Hansen, 1992) or in each of six reviews of

