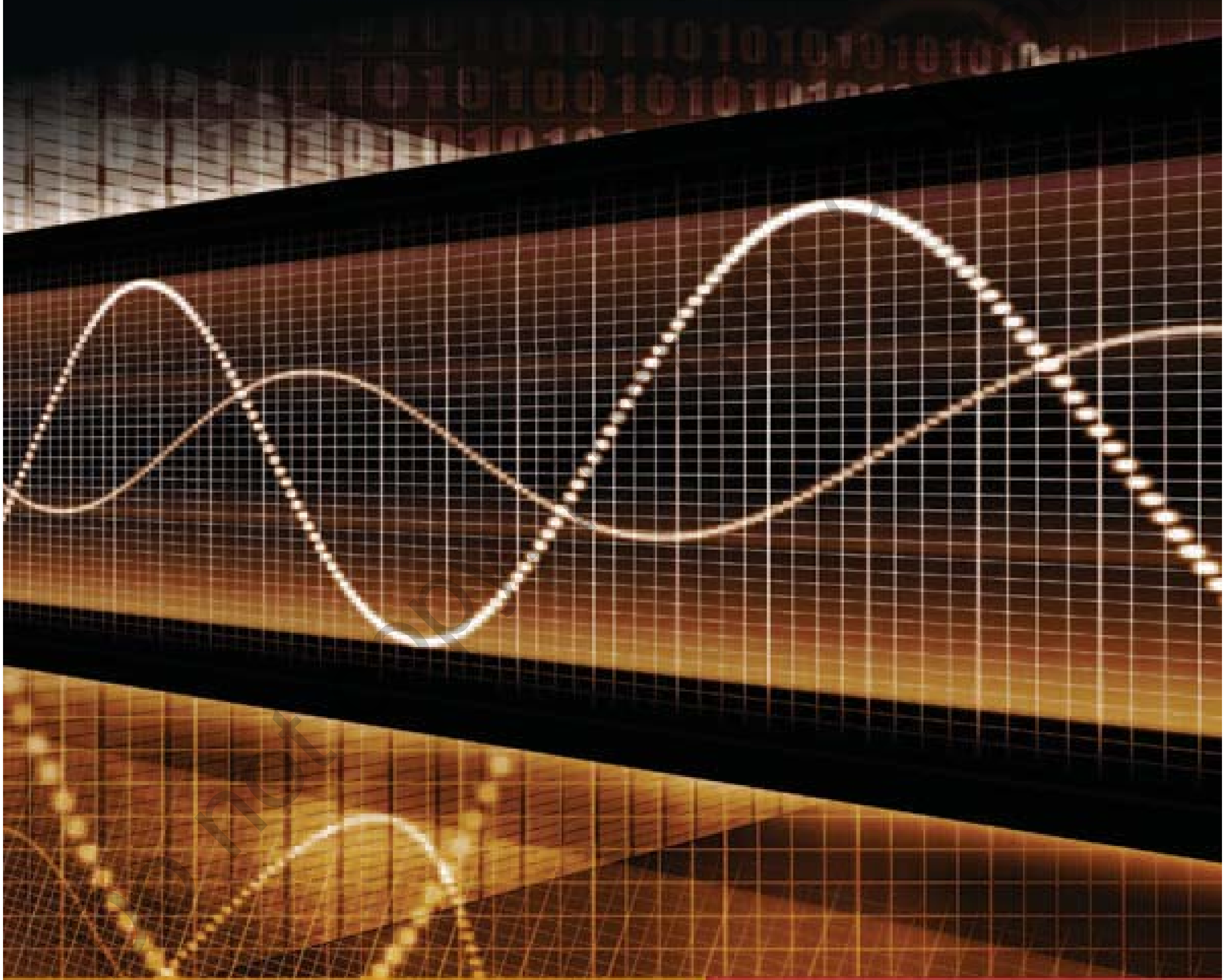
RESEARCH SYNTHESIS AND META-ANALYSIS

A Step-by-Step Approach



HARRIS COOPER

FIFTH EDITION

APPLIED SOCIAL RESEARCH METHODS SERIES

Volume 2

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

6

Step 5

Analyzing and Integrating the Outcomes of Studies

What procedures should be used to condense and combine the research results?

Primary Function in Research Synthesis

To identify and apply procedures for (a) combining results across studies and (b) testing for differences in results between studies

Procedural Variation That Might Produce Differences in Conclusions

Variation in procedures used to summarize and compare results of included studies (e.g., narrative, vote count, averaged effect sizes) can lead to differences in cumulative results.

(Continued)

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

(Continued)

Questions to Ask When Analyzing and Integrating the Results of Studies

1. Was an appropriate method used to combine and compare results across studies?
2. If a meta-analysis was performed, was an appropriate effect size metric used?
3. If a meta-analysis was performed, (a) were average effect sizes and confidence intervals reported and (b) was an appropriate model used to estimate the independent effects and the error in effect sizes?
4. If a meta-analysis was performed, was the homogeneity of effect sizes tested?
5. Were (a) study design and implementation features along with (b) other critical features of studies, including historical, theoretical, and practical variables, tested as potential moderators of study outcomes?

This chapter describes

- A rationale for the use of meta-analyses
- Statistical methods used to summarize research results including
    - Counting study outcomes
    - Averaging effect sizes
    - Examining the variability in effect sizes across studies
- Some practical issues in the application of meta-analytic procedures
- Some advanced meta-analytic procedures

Data analysis involves reducing the separate data points collected by the inquirer into a unified statement about the research problem. It involves ordering, categorizing, and summarizing the data, as well as performing inference tests that attempt to relate data samples to the populations they arise from. Inferences made from data analysis require that decision rules be used to distinguish systematic data patterns from noise (or chance fluctuation). Although different decision rules can be used, the rules involve assumptions about what the target population looks like (e.g., it is normally distributed) and

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

what criteria (e.g., the threshold probability for declaring a finding statistically significant) must be met before an existing pattern in the data is said to be reliable. The purpose of data analysis is to summarize and describe the data in a form that permits valid interpretation.

DATA ANALYSIS IN PRIMARY RESEARCH AND RESEARCH SYNTHESIS

Just as any scientific inquiry requires the leap from concrete operations to abstract concepts, both primary researchers and research synthesists must leap from patterns found in samples of data to more-general conclusions about whether these patterns also exist in the target populations. However, until the mid-1970s, there had been almost no similarity in the analysis techniques used by primary researchers and research synthesists. Primary researchers were obligated to present sample statistics and to substantiate any inferences drawn from their data by providing the results of statistical tests. Most frequently, primary researchers (a) compared sampled means to one another or calculated other measures of relationship, (b) made the assumptions needed for conducting inference tests relating the sample results to populations, and (c) reported the probabilities associated with whether systematic differences in the sample could be inferred to hold in the target population as well.

Traditional statistical aids to primary data interpretation have not gone uncriticized. Some have argued that significance tests are not very informative since they tell only what the likelihood is of obtaining the observed results when the null hypothesis is true (e.g., Cohen, 1994; Cumming, 2012). These critics argue that in a population of people, the null hypothesis is rarely if ever true and therefore the significance of a given test is mainly influenced by how many participants have been sampled. Also, critics who are skeptical about the value of null hypothesis significance testing point to limitations in the generalization of these findings to the target population. No matter how statistically significant a relation may be, the results of a study are generalizable only to people like those who participated in that particular research effort.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Skepticism about the value of statistics helps those who use them refine their procedures and keep their output in proper perspective. Nonetheless, most primary researchers use statistics and most would feel extremely uncomfortable about summarizing the results of their studies without some assistance (or credibility) supplied by statistical procedures. Saying, "I looked at the group means and they looked different to me" is simply not acceptable in primary research.

In contrast to primary researchers, until recently research synthesists were not obligated to apply any statistical techniques in the interpretation of cumulative results. Traditionally, synthesists interpreted data using intuitive rules of inference unknown even to themselves. Analysis methods were idiosyncratic to the perspective of that particular synthesist. Therefore, a description of the common rules of inference used in research syntheses was not possible.

The subjectivity in analysis of research literatures led to skepticism about the conclusions of many syntheses. To address the problem, methodologists introduced quantitative methods into the synthesis process. The methods use the statistics contained in the individual studies as the primary data for the research synthesis.

META-ANALYSIS

I suggested in Chapter 1 that the two events that had the greatest influence on state-of-the-art research synthesis are the growth in the amount of research and the rapid advances in computerized research retrieval systems. A third major influence is the introduction of quantitative procedures, called *meta-analysis,* into the research synthesis process.

The explosion in social science research focused considerable attention on the lack of standardization in how synthesists arrived at general conclusions from series of related studies. For many topic areas, a separate verbal description of each relevant study was no longer possible. One traditional strategy was to focus on one or two studies chosen from dozens or hundreds. This strategy failed to portray accurately the accumulated state of knowledge. Certainly, in

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

areas where dozens or hundreds of studies exist, synthesists must describe prototype studies so that readers understand the methods used by primary researchers.

However, relying on the results of prototype studies to represent the results of all studies may be seriously misleading. First, as we have seen, this type of selective attention is open to confirmatory bias: synthesists may highlight only studies that support their initial position. Second, selective attention to only a portion of all studies places little or imprecise weight on the volume of available tests. Presenting one or two studies without a cumulative analysis of the entire set of results gives the reader no estimate of the confidence that should be placed in a conclusion. Finally, selectively attending to evidence cannot give a good estimate of the strength of a relationship. As evidence on a topic accumulates, researchers become more interested in *how much* of a relationship exists between variables rather than simply *whether a relationship exists at all.*

Synthesists not employing meta-analysis also face problems when they consider the variation between the results of different studies. They will find distributions of results for studies sharing a particular procedural characteristic but varying on many other characteristics. Without meta-analysis, it is difficult to conclude accurately whether a procedural variation affected study outcomes; the variability in results obtained by any single method likely will overlap with the distributions of results of studies using a different method.

It seems, then, that there are many situations in which synthesists need to turn to meta-analytic techniques. The application of quantitative inference procedures to research synthesis was a necessary response to the expanding literature. If statistics are applied appropriately, they should enhance the validity of a synthesis' conclusions. Quantitative research synthesis is an extension of the same rules of inference required for rigorous data analysis in primary research. If primary researchers must specify quantitatively the relation of the data to their conclusions, the next users of the data should be required to do the same. The inference procedure that sounded so ludicrous in the context of a single study ("The means looked different to me") is no less so in the context of research synthesis.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Meta-Analysis Comes of Age

Early on, meta-analysis was not without its critics, and some criticisms persist. Initially, the value of quantitative synthesis was questioned along lines similar to criticisms of primary data analysis (e.g., Barber, 1978; Mansfield & Bussey, 1977). However, much of the criticism stemmed less from issues in meta-analysis than from inappropriate aggregation procedures that are more general, such as a lack of attention to moderating variables, that were incorrectly thought to be caused by the use of quantitative combining procedures when they were really independent (and poor) decisions on the part of the research synthesists. I will return to criticism of meta-analysis, and rigorous research synthesis in general, in the final chapter.

Meta-analysis is now an accepted procedure and its application within the social and medical sciences is on the ascent. Today, literally thousands of meta-analyses have been published, and the number published each year continues to grow larger. Figure 6.1 presents some evidence of this increasing impact in the sciences and social sciences. The figure is based on entries in the Web of Science Core Collection (retrieved April 3, 2015). It charts the growth in the number of documents retrieved by using the topics "research synthesis," "systematic review," "research review," "literature review," and/or "meta-analysis" for even-numbered years from 1996 to 2014. The figure indicates that the total number of references has risen every year without exception and is accelerating. Clearly, the role that research syntheses and meta-analysis play in our knowledge claims is large and growing larger.

When Not to Do a Meta-Analysis

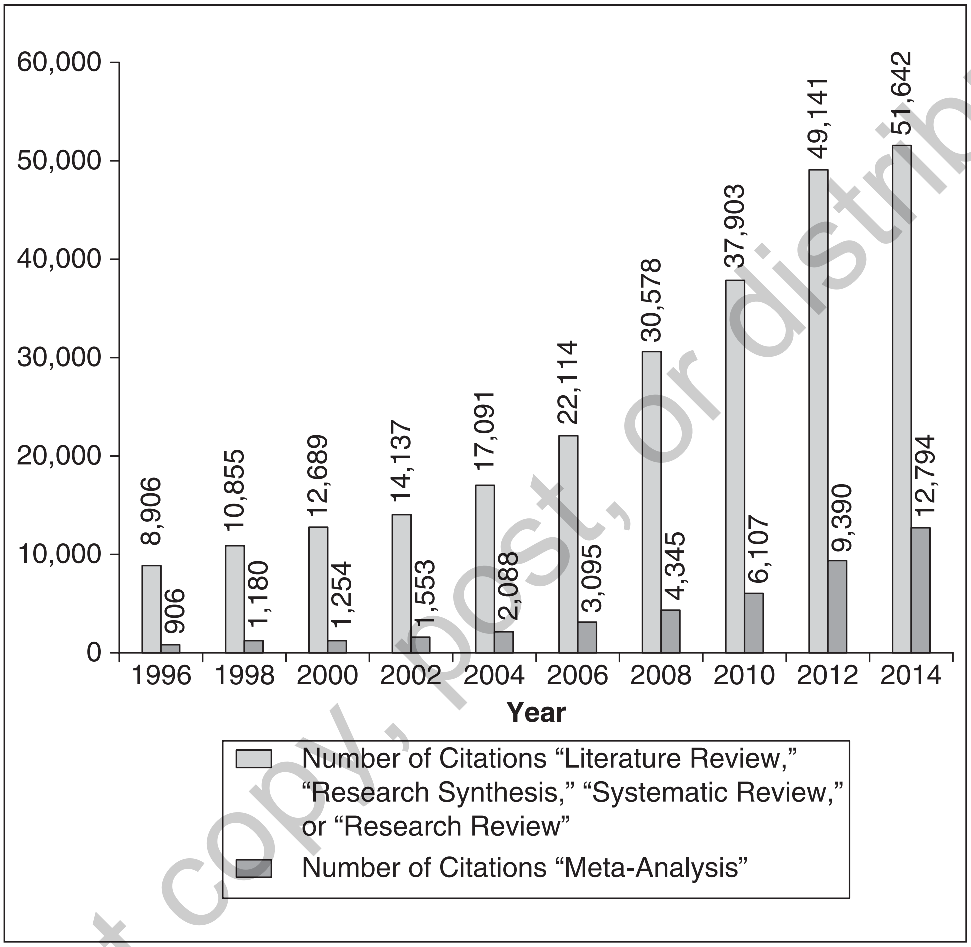
Much of this chapter will describe some basic meta-analysis procedures and how they are applied. However, it is important to state explicitly some circumstances for which the use of quantitative procedures in research syntheses is *not* appropriate.

First, quantitative procedures are applicable only to research syntheses, not to literature reviews with other foci or goals (see Chapter 1).

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Figure 6.1 Web of Science Core Collection Frequency of References to “Research Synthesis,” or “Systematic Review,” or “Research Review,” or “Literature Review,” or “Meta-Analysis”



For instance, if a literature reviewer is interested in tracing the historical development of the concept “intrinsic motivation,” it would not be necessary for him or her to do a quantitative synthesis. However, if the synthesist also intended to make inferences about whether different definitions of intrinsic motivation lead to different research results, then a quantitative summary of relevant research would be appropriate. Also, meta-analysis is not called for if the goal of the literature review is to critically or historically appraise the research study by study or to identify particular studies that are central to a field. In such

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

instances, a proper integration likely would treat the results of studies as an emerging series of events—that is, it would use a historical approach to organizing the literature review rather than a statistical aggregation of the cumulative findings. However, if the synthesists are interested in whether the results of studies *change* over time, then meta-analysis would be appropriate.

Second, the basic premise behind the use of statistics in research syntheses is that a series of studies address an identical conceptual hypothesis. If the premises of a literature review do not include this assertion, then there is no need for cumulative statistics. Related to this point, a synthesist should not quantitatively combine studies at a broader conceptual level than readers would find useful. At an extreme, most social science research could be categorized as examining a single conceptual hypothesis—social stimuli affect human behavior. Indeed, for some purposes such a hypothesis test might be very enlightening. However, the fact that "it can be done" should not be used as an excuse to quantitatively lump together concepts and hypotheses simply because methods are available to do so (see Kazdin, Durac, & Agteros, 1979, for a humorous treatment of this issue). Synthesists must pay attention to those distinctions in the literature that will be meaningful to the users of the synthesis. For example, in the meta-analysis of the effects of choice on intrinsic motivation, we did not combine study results across the nine different outcome measures. Doing so would have obscured important distinctions among the outcomes and might have been misleading. Instead, the highest level of data aggregation was within outcome types.

Another instance of too much aggregation occurs when a hypothesis has been tested using different types of controls. For example, one study examining the effect of daily aerobic exercise on adults' levels of cognitive functioning might compare this treatment to a no-treatment control while another study compares it to a treatment in which participants receive written information about the importance of exercise. It might not be informative to statistically combine the results of these two studies. To what comparison does the combined effect relate? Synthesists might find that a distinction in the type of control group is important enough not to be obscured

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

in a quantitative analysis (but an analysis of the moderating effects of different types of control groups might be appropriate here).

Third, under certain conditions meta-analysis might not lead to the kinds of generalizations the synthesists wish to make. For example, cognitive psychologists or cognitive neuroscientists might argue that their methodologies typically afford good controls and reasonably secure findings because the things they study are not strongly affected by the context in which the study is conducted. Thus, the debate about effects in these areas of research usually occurs with reference to the choice of variables and their theoretical, or interpretive, significance. Under these circumstances, a synthesist might convincingly establish generalization using conceptual and theoretical bridges rather than statistical ones.

Finally, even if synthesists wish to summate statistical results across studies on the same topic, they may discover that only a few studies have been conducted and that these use methodologies, participants, and outcome measures that are decidedly different from one another. In circumstances where multiple methodological distinctions are confounded with one another (e.g., a particular research design occurs very frequently with a particular type of subject), the statistical combination of studies might mask important differences in research that make interpretation of the synthesis findings difficult. In these instances, it may make the most sense not to use meta-analysis, or to conduct several discrete meta-analyses within the same synthesis by combining only those studies that share similar clusters of features.

It is also important to point out that *the use of meta-analysis is no guarantee that the synthesist will be immune from all inferential errors*. The possibility always exists that the meta-analyst has incorrectly inferred a characteristic of the target population. As in the use of statistics in primary research, this can occur because the target population does not conform to the assumptions underlying the analysis techniques or because of the probabilistic nature of statistical findings. If you think that the population statistics do not conform to the assumptions of the statistical test you have chosen, find a more appropriate test or eschew the use of meta-analysis

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

altogether. In sum, then, an important question to ask when evaluating a research synthesis is,

Was an appropriate method used to combine and compare results across studies?

The Impact of Integrating Techniques on Synthesis Outcomes

In Chapter 1 I described a study I conducted with Robert Rosenthal (Cooper & Rosenthal, 1980) in which we demonstrated some of the differences in conclusions that might be drawn by nonquantitative synthesists and meta-analysts. In that study, graduate students and university faculty members were asked to evaluate the same set of studies, but half used quantitative procedures and half used whatever criteria appealed to them. We found that the meta-analysts thought there was more support for the hypothesis and a larger relationship between variables than did the non-meta-analysts. Meta-analysts also tended to view future replications as less necessary than did non-meta-analysts, although this finding did not reach statistical significance.

It is also likely that the different statistical procedures used by meta-analysts will create variance in synthesis conclusions. Several different paradigms have emerged for quantitatively integrating research with a traditional inference testing model (Hedges & Olkin, 1985; Rosenthal, 1984; Schmidt & Hunter, 2015), while others use a Bayesian perspective (Sutton, Abrams, Jones, Sheldon, & Song, 2000; United States Department of Health and Human Services Agency for Healthcare Research and Quality, 2013). The different techniques generate different output. Thus, the rules adopted to carry out quantitative analysis can differ from synthesist to synthesist, which might create differences in how synthesis results are interpreted. We can assume as well that the rules used by nonquantitative synthesists also vary, but that because of their inexplicit nature it is difficult to compare them formally.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

MAIN EFFECTS AND INTERACTIONS IN META-ANALYSIS

Before examining several of the quantitative techniques available to synthesists, it is important to take a closer look at some of the unique features of accumulated research results. In Chapter 2 on problem formulation, I pointed out that most research syntheses first focus on tests of main effects that were carried out in the primary studies. This is largely because conceptually related replications of main effects occur more frequently than tests of three or more interacting variables. So, for example, you are likely to find in primary studies many more main-effect tests of whether choice influences intrinsic motivation than you are to find tests of interactions of whether this relationship is influenced by the number of choices given. Keep in mind that I am referring here to interaction tests within a single study, not your ability to test for the influence of number-of-choices at the synthesis level because different studies have varied in the number of choices they provide in their test of the main effect.

It is not that interactions tested in primary studies cannot be combined. However, such replications are fewer and, we shall see in the next chapter, their interpretation can be a bit more complex. There are two different ways that interactions tested in primary research could be statistically combined across studies. First, the relationship strengths associated with each study's interaction test could be aggregated. An alternative strategy would be to aggregate separately the relationship of two of the interacting variables at each level of the third variable. For instance, assume there exists a set of studies in which the primary researchers tested whether the effect of choice in intrinsic motivation differed depending on the number of choices given to participants. The synthesists could generate an estimate of the difference in intrinsic motivation depending on the number of choices given. They could aggregate all motivation measures taken under conditions where a choice between two alternatives was compared to no choice. They could do the same for measures taken after, say, two or three choices. Then, the different effect sizes could be compared. This would probably be

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

more useful and easily interpretable than a direct estimate of the magnitude of the interaction effect. However, in order to do this, the primary research reports must contain the information needed to isolate the different simple main effects. The synthesist might also have to group numbers of choices (e.g., three to five choices and six or more choices) in order to have enough tests to generate a good estimate.

Because main effects are most often the focus of meta-analysts and in many instances meta-analysts interested in interactions reduce them to simple effects, my discussion of the quantitative combining techniques will refer to main effects only. The generalization to meta-analyzing interactions is mathematically straightforward.

META-ANALYSIS AND THE VARIATION AMONG STUDY RESULTS

In research syntheses, the most obvious feature of both main effects and interactions is that the results of the separate tests of the same relationship will vary from one study to the next. This variability is sometimes dramatic and requires us to ask where the variability comes from.

Sources of Variability in Research Findings

Differences in the outcomes of studies can be caused by two types of influences. The simplest cause is the one that is most often overlooked by nonquantitative synthesists—sampling variability. Even before the current interest in quantitative synthesis, Taveggia (1974) recognized this important influence:

> A methodological principle overlooked by writers of . . . reviews is that research results are *probabilistic*. What this principle suggests is that, in and of themselves, the findings of any single research are meaningless—they may have occurred simply by chance. It also follows that if a large enough number of researches has been done on a particular topic, chance alone dictates that studies will exist

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

> that report inconsistent and contradictory findings! Thus, what appears to be contradictory may simply be the positive and negative details of a distribution of findings. (pp. 397–398, emphasis in original)

Taveggia highlights one of the implications of using probability theory and sampling techniques to make inferences about populations.

As an example, suppose it was possible to measure the academic achievement of every American student as well as whether each student did homework. Also, suppose that if such a task were undertaken, it would be found that achievement was exactly equal for students who do and do not do homework—that is, exactly equal achievement test mean scores existed for the two subpopulations. Still, if 1,000 samples of 50 homeworkers and 50 no-homeworkers were drawn from this population, very few comparisons between samples would reveal exactly equal group means. About half would show homeworkers achieving better and half would show no-homeworkers achieving better. Furthermore, if the sample means were compared statistically using a *t*-test and the $p < .05$ significance level (two-tailed), about 25 comparisons would show a significant difference favoring homeworkers while about 25 would favor no-homeworkers. This variation in results is an unavoidable consequence of the fact that the means estimated by sampling will vary somewhat from the true population values. And, just by chance alone, some comparisons will pair sample estimates that vary from their true population values by large amounts and in opposite directions.

In the example given, it is unlikely that you would be fooled into thinking anything but chance caused the result—after all, 950 comparisons would reveal nonsignificant differences and significant results would be distributed equally for both significant positive and negative outcomes. However, in practice the pattern of results is rarely this clear. As we discovered in the chapter on literature searching, you might not be aware of all null results because they are hard to find. Complicating matters further, even if an overall relation does exist between two variables (i.e., the null hypothesis is false), some studies can still show significant results in a direction opposite to the relation in the population.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

To continue the example, if the average achievement of homeworkers is better than no-homeworkers, some comparisons of samples randomly drawn from the two subpopulations will still favor no-homeworkers, the number depending on the size of the relation, the size of the samples, and how many comparisons have been performed. In sum, then, one source of variance in the results of studies can be chance fluctuations due to the inexactness of estimates based on samples drawn from populations.

A second source of variance in study outcomes is of more interest to synthesists. This variance in results is created by differences in how studies are conducted. This variance is added to the variance due to sampling participants. Just as people are sampled, you can think of a set of studies as a sample of studies drawn from a population of all possible studies. And, because studies can be conducted in different ways (just as people can differ in personal attributes) that affect the studies outcomes, a sample of studies also will exhibit chance variation from other possible samples of studies. For instance, the homework synthesists might find that studies comparing achievement among students who do and do not do homework have been conducted with students at different grade levels; with unit tests, class grades, or standardized tests as measures of achievement; and with an assortment of classes with different subject matters. Each of these differences in the studies' methods or contexts could create variation in study results and therefore could create results that differ randomly from another sample of studies drawn from the same population of studies. This variation will be added to the variation caused by the sampling of study participants from the population of participants.

It is also possible that this variation associated with study-level differences is systematically related to the variation in study results. For example, homework studies conducted with elementary school students might produce results that differ systematically from studies conducted with high school students. In Chapter 2, the notion of synthesis-generated evidence was introduced to describe what we learn when we find associations between study characteristics and study outcomes.

The existence of the two sources of variance in research results—the one generated by sampling participants and the other by sampling

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

studies—raises an interesting dilemma. When discrepant findings occur within a set of studies, should you seek an explanation for them by attempting to identify systematic differences in results associated with differences in the methods used in studies? Or should you simply assume the discrepant findings were produced by variations due to sampling (of participants and/or study procedures)? Some tests have been devised to help you answer this question. In effect, these tests use sampling error (associated with participants or both participants and studies) as the null hypothesis. They estimate the amount of variance in findings that would be expected if sampling error alone were making the study findings different.[1] If the observed variation in results across studies is too great to be explained by sampling error alone, then the null hypothesis is rejected. It suggests that the notion that all the results were drawn from the same population of results can be rejected.

In the sections that follow, I will introduce some of the quantitative synthesis techniques that are available to you. I have chosen the techniques because they are relatively simple and broadly applicable. The treatment of each technique will be conceptual and introductory but detailed enough to permit you to perform a sound, if basic, meta-analysis. You can consult the primary sources cited in the text if (a) you want a more detailed description of these techniques and their variations, including how they are derived, and/or (b) your meta-analysis has some unique possibilities for exploring data in ways not covered here. For the discussion that follows, I have assumed you have a working knowledge of the basic inferential statistics employed in the social sciences.

Before I begin, though, there are three assumptions crucial to the validity of a conclusion based on an integration of statistical findings from individual studies. First and most obviously, *the individual findings that go into a cumulative analysis should all test the same comparison or estimate the same relationship*. Regardless of how conceptually broad or narrow your ideas might be, you should be comfortable with the assertion that the included statistical tests from the primary studies address the same question. Second, *the separate tests that go into the cumulative analysis must be independent of one another.* Identifying independent comparisons was discussed in Chapter 4, on gathering

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

information from studies. You must take care to identify comparisons so that each one contains unique information about the hypothesis. Finally, you must believe that *the primary researchers made valid assumptions when they computed the results of their tests*. Thus, for example, if you want to combine the effect sizes resulting from comparisons between two means, you must assume that the observations in the two groups in the primary studies are independent and normally distributed, and that their variances are roughly equal to one another.

VOTE COUNTING

The simplest methods for combining independent statistical tests are the vote counting methods. Vote counts can take into account the statistical significance of findings or focus only on the direction of the findings.

For the first method, the meta-analysts would take each finding[2] and place it into one of three categories: statistically significant findings in the expected direction (I will refer to these as positive findings), statistically significant findings in the unexpected (negative) direction, and nonsignificant findings—that is, findings that did not permit rejection of the null hypothesis. The meta-analysts then might establish the rule that the category with the largest number of findings tells what the direction of the relationship is in the target population.

This vote count of significant findings has much intuitive appeal and has been used quite often. However, the strategy is unacceptably conservative and often can lead to erroneous conclusions (Hedges & Olkin, 1980). The problem is that using the traditional definition of statistical significance, chance alone should produce only about 5% of all findings falsely indicating a significant effect. Therefore, much fewer than one-third positive and statistically significant findings might indicate a real difference exists in the target population. This vote-counting strategy requires that at least 34% of findings be positive and statistically significant before a result is declared a winner.

Let me illustrate just how conservative this approach is. Assume that a correlation of $r = .30$ exists between two variables in a population

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

and 20 studies have been conducted with 40 people in each sample (this would not be an uncommon scenario in the social sciences). The probability that the vote count associated with this series of studies will conclude a positive relation exists—if the plurality decision rule described in the preceding paragraph is used—is less than 6 in 100. Thus, the vote count of significant findings could, and often does, lead vote counters to suggest accepting the null hypothesis, and perhaps abandoning fruitful theories or effective interventions when, in fact, no such conclusion is warranted.

Adjusting the frequencies of the three types of findings (positive, negative, and null) so that the true expected percentage of each finding (95% null and 2.5% significant in each direction) is taken into account solves one problem but it highlights another one. We have seen that null results are less likely to be reported by researchers and are less likely to be retrieved by synthesists. Therefore, if the appropriate expected values are used in a vote-count analysis, it could often occur that *both* positive and negative significant findings appear more frequently than would be expected by chance alone. Thus, it seems that using the frequency of nonsignificant findings in a vote count procedure is of dubious value.

An alternative vote-counting method is to compare the frequency of statistically significant positive findings against the frequency of significant negative ones. This procedure assumes that if the null hypothesis prevails in the population, then the frequency of significant positive and negative findings is expected to be equal. If the frequency of findings is found not to be equal, then the null hypothesis can be rejected in favor of the prevailing direction. A problem with this vote-count approach is that the expected number of nonsignificant findings, even when the null hypothesis is not true, can still be much greater than the expected number of either positive or negative significant findings. Therefore, this approach will ignore many findings (all nonsignificant ones) and will be very low in statistical power.

A final way to perform vote counts in research synthesis involves tallying the number of positive and negative findings regardless of their statistical significance. In this approach, the meta-analyst categorizes findings based solely on the direction of their outcome,

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

ignoring their statistical significance. Again, if the null hypothesis is true—that is, if no relationship exists between the variables in the sampled population—we would expect the number of findings in each direction to be equal.

Once the number of results in each direction is counted, the meta-analyst can perform a simple sign test to discover if the cumulative result suggests that one direction occurs more frequently than would be expected by chance. The formula for computing the sign test is as follows:

$$Z_{vc} = \frac{(N_p) - (\tfrac{1}{2}N)}{\tfrac{1}{2}\sqrt{N}} \tag{1}$$

where

Z_{vc} = the standard normal deviate, or Z-score, for the overall series of findings;

N_p = the number of positive findings; and

N = the total number of findings (positive plus negative findings).

The Z_{vc} can be referred to a table of standard normal deviates to discover the probability (one-tailed) associated with the cumulative set of directional findings. If a two-tailed p-level is desired, the tabled p-value should be doubled. The values of Z associated with different p-levels are presented in Table 6.1. This sign test can be used in a vote count of either the simple direction of all findings or the direction of only significant findings, though using the direction of findings is recommended.

Suppose 25 of 36 comparisons find that adults given an intervention to increase aerobic activity exhibited better neurocognitive functioning than those in a no-intervention group. The probability that this many findings would be in one direction, given that in the target population (of all intervention tests) there is equal neurocognitive functioning exhibited by people in the two conditions, is $p < .02$ (two-tailed) associated with a Z_{vc} of 2.33. This result would lead the meta-analyst to conclude a positive intervention effect was supported by the series of findings.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The vote-count method that uses the direction of findings regardless of significance has the advantage of using information from all statistical findings. Still, it has some drawbacks. Similar to the other vote-count methods, it does not weight a finding's contribution to the overall result by its sample size. Thus, a finding based on 100 participants is given weight equal to one with 1,000 participants. Furthermore, the revealed magnitude of the relationship (e.g., the impact of the treatment) in each finding is not considered—a finding showing a large increase in cognitive functioning due to the intervention is given equal weight to one showing a small decrease in functioning. Finally, a practical problem with the directional vote count is that primary researchers frequently do not report the direction of findings if a comparison proved statistically nonsignificant.

Still, the vote count of directional findings can be an informative complement to other meta-analytic procedures, and can even be used to generate an estimate of the strength of a relationship. Bushman and Wang

Table 6.1 Standard Normal Deviation Distribution

z-score	Area *z* to –*z*	*p*-level 2-tailed	*p*-level 1-tailed
2.807	.995	.005	.0025
2.576	.99	.01	.005
2.432	.985	.015	.0075
2.326	.98	.02	.01
2.241	.975	.025	.0125
2.170	.97	.03	.015
2.108	.965	.035	.0175
2.054	.96	.04	.02
2.000	.954	.046	.023
1.960	.95	.05	.025
1.881	.94	.06	.03
1.751	.92	.08	.04
1.645	.9	.1	.05
1.440	.85	.15	.075
1.282	.8	.2	.10
1.150	.75	.25	.125
1.036	.7	.3	.150
0.842	.6	.4	.20
0.674	.5	.5	.25
0.524	.4	.6	.30
0.385	.3	.7	.35
0.253	.2	.8	.40
0.126	.1	.9	.45

SOURCE: Adapted from: Wikipedia (2015), http://en.wikipedia.org/wiki/Standard_normal_table

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

(2009) provide formulas and tables that can be used to estimate the size of a population relationship given that the meta-analysts know (a) the number of findings, (b) the direction of each finding, and (c) the sample size of each finding. For example, let's assume that each one of the 36 comparisons between an activity intervention and no-intervention group was based on a sample size of 50 participants. Using Bushman and Wang's technique, I find that when 25 of the 36 (69%) comparisons revealed better cognitive functioning in the intervention group, the most likely population value for a correlation between group membership and activity is $r = .07$. Of course, this example is artificial because I assumed all the sample sizes were equal. The calculations are more complex in many circumstances, not only because sample sizes vary but also because you will have comparisons (votes) for which you have no direction. This complicates the estimating technique greatly. In the past, when we have used this technique (see Cooper, Charlton, Valentine, & Muhlenbruck, 2000), we conducted the analyses several times, using different sets of assumptions. In general, this technique should be used with caution and only in conjunction with other meta-analytic techniques that produce conclusions that are less tentative.

In sum, then, meta-analysts can perform vote counts to aggregate results across individual studies by comparing the number of directional findings and/or the number of significant directional findings. Both of these procedures will be very imprecise and conservative—that is, they will accept the null hypothesis when more-precise methods suggest it should be rejected. The simple direction of results will not appear in many research reports in the first case, and nonsignificant findings cannot contribute to the analysis in the second case. Vote counts can be described in meta-analyses but should be used to draw inferences only in combination with more sensitive meta-analysis procedures.

Do not copy, post, or distribute

Combining Significance Levels

One way to address the shortcomings of vote counts is to consider combining the exact probabilities associated with the results of each comparison. Rosenthal (1984) cataloged 16 methods for combining the results of inference tests so that an overall test of the null hypothesis

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

can be obtained. By using the exact probabilities, the results of the combined analysis take into account the different sample sizes and relationship strengths found in each comparison. Thus, the combining-significance-levels procedure overcomes the improper weighting problems of the vote count. However, it has severe limitations of its own. First, as with vote counts, the combining-probability procedures answer the "yes or no?" questions but not the "how much?" question. Second, whereas the vote-count procedure is overly conservative, the combining-significance-levels procedure is extremely powerful. In fact, it is so powerful that for hypotheses or relationships that have generated a large number of findings, rejecting the null hypothesis is so likely, because even very small relationships can produce significant combined probabilities, that it becomes a rather uninformative exercise. For this reason, these procedures have largely fallen out of use.

MEASURING RELATIONSHIP STRENGTH

The primary function of the procedures described so far is to help meta-analysts accept or reject the null hypothesis. Until recently, most researchers interested in social theory and the impact of social interventions have been content to simply identify relations that have some explanatory value. The prevalence of this "yes or no" question was partly due to the relatively imprecise nature of social science theories and hypotheses. Social hypotheses typically were crudely stated first approximations to the truth. Social researchers rarely asked how potent theories or interventions were for explaining human behavior or how competing explanations compare with regard to their relative explanatory value. Today, as their theories and interventions are becoming more sophisticated, social scientists are more often making inquiries about the size of relationships.

Giving further impetus to the "how much?" question is a growing disenchantment with the null hypothesis significance test itself. As I noted earlier, whether a null hypothesis can be rejected is tied closely to the particular research project under scrutiny. If an ample number of participants are available or if a sensitive research design is employed, a rejection of the null hypothesis often is not surprising. This

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

state of affairs becomes even more apparent in meta-analyses that include a combined significance level, where the power is great to detect even very small relations. A null hypothesis rejection, then, does not guarantee that an important social insight has been achieved.

Finally, when used in applied social research, the vote-count and combined-significance-level techniques give no information on whether the effect of a treatment or the relationship between variables is large or small, important or trivial. For example, if we find that the relationship between whether a participant (a) is an adolescent or adult and (b) believes that women share some culpability when a rape occurs is statistically significant and the correlation is $r = .01$, is this a strong enough relationship that it should influence how interventions are delivered? What if the result is statistically significant and the correlation is $r = .30$? This example suggests that the "yes or no?" question is often not the question of greatest importance. Instead, the important question is, "How much does the age of the participant influence beliefs about rape?" The answer might be zero or it might suggest a small or large relationship. The answer to this question could help meta-analysts (and others) make recommendations about how best to construct rape-attitude interventions so they are most effective. Given these questions, meta-analysts would turn to the calculation of average effect sizes. Also, as we shall see shortly, the null hypothesis question, "Is the relationship different from zero?" can be answered by placing a confidence interval around the "how much?" estimate, removing the need for separate null hypothesis significance tests.

Definition of Effect Size

In order to answer meaningfully the "how much?" question, we must agree on definitions for the terms *magnitude of difference, relationship strength,* or what generally is called the *effect size.* Also, we need methods for quantitatively expressing these ideas once we have defined them. Jacob Cohen's (1988) book *Statistical Power Analysis for the Behavioral Sciences* presented what is now the standard definition of effect sizes. He defined an effect size as follows:

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

> Without intending any necessary implication of causality, it is convenient to use the phrase "effect size" to mean "*the degree* to which the phenomenon is present in the population," or "the degree to which the null hypothesis is false." By the above route it can now readily be clear that when the null hypothesis is false, it is false to some specific degree, i.e., *the effect size (ES) is some specific non-zero value in the population*. The larger this value, the greater the *degree* to which the phenomenon under study is manifested. (pp. 9–10, emphasis in original)

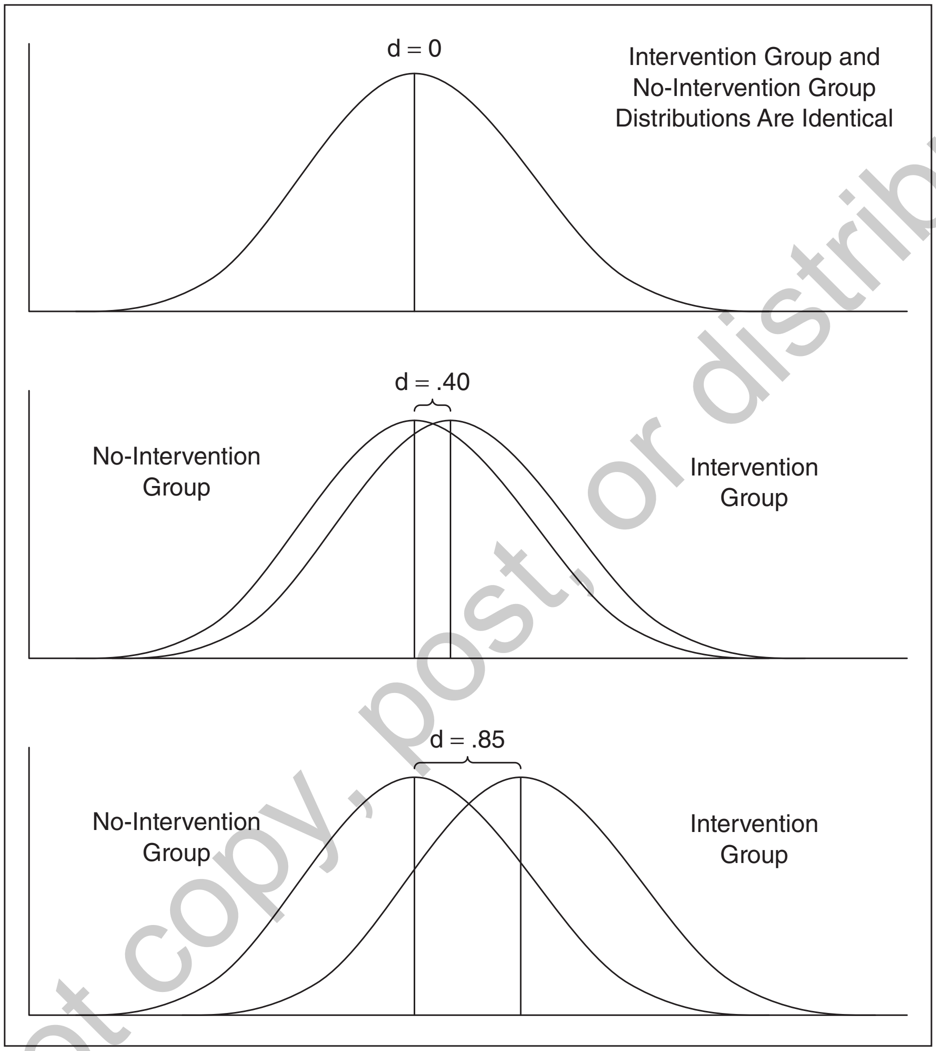
Figure 6.2 presents three hypothetical relationships that illustrate Cohen's definition. Suppose the results come from three experiments comparing the effects of an aerobic exercise intervention versus a no-treatment control on adults' cognitive functioning. The top graph presents a null relationship. That is, the participants given the intervention have a mean and distribution of cognitive functioning scores identical to the no-intervention participants. In the middle graph, the intervention group has a mean cognitive functioning score slightly higher than that of the no-intervention group, and in the bottom graph the difference between intervention and no-intervention is even greater. A measure of effect size must express the three results so that greater departures from the null are associated with larger effect size values.

Cohen's (1988) book contains many different metrics for describing the strength of a relationship. Each effect size index is associated with a particular research design in a manner similar to *t*-tests being associated with two-group comparisons, *F*-tests associated with multiple-group designs, and chi-squares associated with frequency tables. Next, I will describe the three primary metrics used by the vast majority of meta-analysts. These metrics are generally useful—almost any research outcome can be expressed using one of them. For more-detailed information on these effect size metrics, as well as many others, the reader should consult Cohen's (1988) book or Cumming's (2012) book. However, Cohen describes several metrics that permit effect size estimates for multiple-degree-of-freedom comparisons (e.g., a comparison involving more than two group means, such as three religious groups' attitudes toward rape), and these typically should not be used, for reasons that will be discussed

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Figure 6.2 Three Hypothetical Relations Between an Exercise Intervention and a No-Intervention Group



shortly. Thus, my description of metrics is restricted to those commensurate with single-degree-of-freedom tests.

Standardized Mean Difference: The *d*-index or *g*-index

The *d*-index, or standardized mean difference measure, of an effect size is appropriate to use when the difference between two means is

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

being compared. The *d*-index is typically used in association with *t*-tests or *F*-tests based on a comparison of two groups or experimental conditions. The *d*-index expresses the distance between the two group means in terms of their common standard deviation. By the term *common standard deviation,* I mean that the assumption is made that if we could measure the standard deviations within the two subpopulations sampled into the two groups, we would find them to be equal.

The hypothetical research results for three studies presented in Figure 6.2 comparing an intervention meant to promote aerobic activity among adults with a no-intervention condition illustrates the *d*-index. The dependent variable is some measure of neurocognitive functioning, maybe short-term memory or speed of processing. For the top graph, the research result supports the null hypothesis and the *d*-index equals zero. That is, there is no distance between the means of the exercise intervention and no-intervention group. The middle research result reveals a *d*-index of .40—that is, the mean of the intervention group lies 4/10ths of a standard deviation to the right of the no-intervention group's mean. In the third example, a *d*-index of .85 is portrayed. Here, the intervention group mean rests 85/100ths of a standard deviation to the right of the mean of the no-intervention group.

Calculating the *d*-index is simple. The formula is as follows:

$$d = \frac{\overline{X_1} - \overline{X_2}}{SD_{\text{within}}} \qquad (2)$$

where

$\bar{X}_1$ and $\bar{X}_2$ = the two group means; and

SD_{within} = the estimated common standard deviation of the two groups.

To estimate SD_{within}, you can use the formula

$$SD_{\text{within}} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}} \qquad (3)$$

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

where

SD_1 and SD_2 = the standard deviations of Group X_1 and Group X_2, respectively, and

n_1 and n_2 = the sample sizes in Group X_1 and Group X_2, respectively.

The *d*-index is not only simple to compute, but is also scale free. That is, the standard deviation adjustment in the denominator of the formula means that studies using different measurement scales can be compared or combined. So, for example, if one study of the exercise intervention's effect used a measure of short-term memory as the outcome measure and another study used a measure of processing speed as the outcome measure, it would make little sense to combine the two raw differences between the intervention and no-intervention group means—that is, combine the numerators of the *d*-index formula. However, it might make sense to combine the two results if we first convert each to a standardized mean difference. Then, if we assume the two outcomes measure the same underlying conceptual variable (i.e., cognitive functioning), the two outcomes have been transformed to a common metric.

The variance of the *d*-index can be closely approximated using the following formula:

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)} \qquad (4)$$

where

all variables are defined as above.

The 95% confidence interval for the *d*-index is then computed as $d - 1.95\,\sqrt{\mathrm{v}_\mathrm{d}} \leq d \geq d + 1.95\,\sqrt{\mathrm{v}_\mathrm{d}}$.

In many instances, meta-analysts will find that primary researchers do not report the means, standard deviations, and sample sizes of the separate groups but do report the *t*-test or *F*-test associated with the difference in means, and the direction of their relationship. In such cases, Rosenthal (1984) provided a computation formula

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

that closely approximates the *d*-index and does not require the meta-analysts to have specific means and standard deviations. This formula is as follows:

$$d = \frac{2t}{\sqrt{df_{\text{error}}}} \tag{5}$$

where

t = the value of the *t*-test for the associated comparison, and

df_{error} = the degrees of freedom associated with the error term of the *t*-test ($n_1 + n_2 - 2$).

In instances where *F*-tests with a single degree of freedom in the numerator are reported, the square root of the *F*-value (i.e., $t = \sqrt{F}$) and its denominator degrees of freedom can be substituted in the above formula. Again, these approximations of the *d*-index assume the meta-analysts know the direction of the mean difference.

In fact, it is possible to calculate *d*-indexes from lots of different pieces of data and from numerous different designs. I refer you to the *Practical Meta-Analysis Effect Size Calculator* (Wilson, 2015). This free website will calculate the *d*-index for you based on 30 variations in the information you have available and for different research designs. Some meta-analysis software programs will also calculate effect sizes for you but you must be sure the available options match the type of data and design you are working with. If not, you can calculate the effect size using an (reliable) Internet calculator and transfer these to the meta-analysis program.

Removing small sample bias from estimates of population values: The g-index. A sample statistic—be it an effect size, a mean, or a standard deviation—typically is based on measurements taken on a small number of people drawn from a larger population. These sample statistics will differ in known ways from the values obtained if we could measure every person in the population. Meta-analysts have devised ways to adjust for the known biases that occur because effect size estimates based on samples are not always unbiased reflections of their underlying population values.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Hedges (1980) showed that the *d*-index based on small samples may slightly overestimate the size of an effect in the population. However, the bias is minimal if the sample size is more than 20. If meta-analysts are calculating standardized mean differences from primary research based on samples smaller than 20, Hedges' *g*-index should be used. The difference between the *d* and *g* formulas is simply that in the *g*-index formula the pooled estimate for the population standard deviation is substituted for the pooled sample standard deviation in the denominator of Formula (2). Conveniently, a search of the Internet for "Effect Size Calculators" will locate websites that will simultaneously calculate for you effect size estimates based on several different formulas (e.g., Ellis, 2009).

In addition to the small sample bias in effect size estimates meta-analysts should always be cautious in interpreting any statistics based on a small number of data points. When samples are small, a single extreme value can create an exceptionally large effect size estimate.

Choosing an estimate for the standard deviation of the d-*index*. Clearly, an important influence on the *d*-index is the size of the standard deviation used to estimate the variance around group means. I mentioned previously that the *d*-index formula is based on the assumption that the standard deviations would be equal in the two groups if they could be measured precisely. Many times, meta-analysts have no choice but to make this assumption because the *d*-index must be estimated from an associated *t*-test or *F*-test, which also makes this assumption. However, in instances where information about standard deviations is available and they appear to be unequal, the meta-analyst can choose one group's standard deviation to serve as the denominator in the *d*-index for purposes of standardizing the mean difference. For example, if an intervention and no-intervention group are being compared and the standard deviations appear to be different (perhaps because the intervention shifts the group mean and also creates greater variance in outcomes), then the control group standard deviation should be used.

Effect Sizes Based on Two Continuous Variables: The *r*-Index

A second effect size, the *r*-index, is simply the Pearson product-moment correlation coefficient. The *r*-index is the most appropriate metric for

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

expressing an effect size when the researcher is interested in describing the relationship between two continuous variables. So, for example, if we are interested in the relationship between participants' amount of exposure to pornography and their degree of belief that women share culpability for rape, we would use the correlation coefficient to estimate this association.

The *r*-index is familiar to most social scientists but the formula for it requires both the variances and covariances of the two continuous variables, so it rarely can be computed from information typically presented in primary research reports. Luckily, primary researchers do report their *r*-indexes in most instances where they are applicable. However, if only the value of the *t*-test associated with the *r*-index is given, the *r*-index can be calculated using the following formula:

$$r = \sqrt{\frac{t^2}{t^2 + df_{\text{error}}}} \tag{6}$$

where

all terms are defined as above.

The variance of the *r*-index can be calculated using the following formula:

$$v_r = \frac{(1 - r^2)^2}{n - 1} \tag{7}$$

where

all terms are defined as above.

The formula can be used to calculate the 95% confidence interval as $r - 1.95\sqrt{v_r} \le r \ge r + 1.95\sqrt{v_r}$.

Normalizing the distribution of r*-indexes.* When *r*-indexes are large—that is, when they estimate population values very different from zero—they will exhibit non-normal sampling distributions. This occurs because *r*-indexes are limited to values between +1.00 and −1.00. Therefore, as a population value approaches either of these limits, the

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

range of possible values for a sample estimate will be restricted on the tail toward the approached limit (see Shadish & Haddock, 2009).

To adjust for this, most meta-analysts convert *r*-indexes to their associated *z*-scores before the effect size estimates are combined or tested for moderators. The *z*-scores have no limiting value and are normally distributed. Conceptually, the transformation "stretches" the restricted tail of the distribution and restores the bell shape of the curve. Once an average *z*-score has been calculated, it can be converted back to an *r*-index. An examination of *r*-to-*z* transformations reveals that the two values are nearly identical until the absolute value of *r* equals about .25. However, when the *r*-index equals .50, the associated *z*-score equals .55, and when the *r*-index equals .8, the associated *z*-score equals 1.1. The *z*-score can also be calculated directly from

$$z = .5\ [\ln(1+\mathrm{r}) - \ln(1-\mathrm{r})] \tag{8}$$

where

ln = natural logarithm and

all other terms are defined as above.

The variance of the *z*-score is

$$v_z = \frac{1}{(n-3)} \tag{9}$$

where

all terms are defined as above.

For greatest ease, you can find *r*-to-*z* transform calculators on the Internet (e.g., http://vassarstats.net/tabs_rz.html) that will also calculate measures of dispersion. Be sure to remember that once you have calculated the average *z*-score of the transformed correlations, you must transform this back into a correlation coefficient when you present your results. The *z*-score will have little meaning for your audience.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Effect Sizes Based on Two Dichotomous Variables: The Odds and Risk Ratios

A third class of effect size metric is applicable when both variables are dichotomous—for example, when elderly adults either receive or do not receive an aerobic activity treatment and the outcome variable is whether or not they are diagnosed with Alzheimer's disease five years later. In this case, one measure of effect, called an *odds ratio*, is often used in medical research, where researchers are frequently interested in the effect of a treatment on mortality or the appearance or disappearance of disease. It is used also in criminal justice research where the outcome variable might be recidivism (re-arrest after the passage of a certain amount of time) or in education studies—for example, when high school graduation (yes or no) is the outcome of interest.

As its name implies, the odds ratio describes the relationship between two sets of odds. For example, suppose meta-analysts come across a study of the effects of an intervention promoting aerobic exercise among elderly adults. Two hundred randomly assigned participants either received or did not receive the intervention; 5 years later they were assessed for the presence of Alzheimer's disease. The results of the study were as follows:

	Intervention	**No Intervention**
No Alzheimer's Disease Indicated	75	60
Alzheimer's Indicated	25	40

In order to calculate an odds ratio, the meta-analysts first determine that the odds against a participant in the intervention condition having Alzheimer's disease were 3 to 1 (75 to 25). The odds against having Alzheimer's disease in the no-intervention condition were 1.5 to 1 (60 to 40). In this case, the odds ratio is 2, meaning the odds of finding evidence of the disease in the no-intervention group were twice those in the intervention group. When the odds are the same in both conditions (i.e., when the treatment had no effect or the null hypothesis was true),

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

the odds ratio will be 1. The odds ratio can be calculated directly from the table by dividing the product of the main diagonal elements by the product of the off-diagonal elements, in our example (75 × 40)/(60 × 25).

Another measure of effect for two dichotomous variables is the risk ratio. This expresses the relative risk of one condition against the other. So, in the example about the risk of getting Alzheimer's disease among the elderly adults who received the intervention was .25, or 25 chances in 100. For no intervention, the risk was .40, or 40 in 100. The risk ratio is then the ratio of these two numbers: .625 if the treated condition is in the numerator or 1.60 if the untreated condition is in the numerator.

Again, the *Practical Meta-Analysis Effect Size Calculator* (Wilson, 2015) can calculate both odds ratios and risk ratios for you. Similar to the *r*-index, before you calculate an average ratio, the individual ratios should be transformed to their log (also provided by the calculator). Then, the average should be transformed back for purposes of interpretation.

Because the odds ratio is used less often in the social sciences, it will not be treated extensively in the next section. However, most of the techniques discussed in the next section are easily adapted to its use. There are many other metrics that can be used when two dichotomous variables are being related to one another; Fleiss and Berlin (2009) provide an overview of numerous effect size estimates gauging the relationship between two dichotomous variables.

As general rules, I have two suggestions when you use effect size calculators available on the Internet. First, check the formulas used in these programs. They might differ in some ways from my simple formulas given above. As long as the website comes from a reliable source, the calculations should be reliable but it is always good to calculate a few effect sizes by hand. This way you can be more confident you understand how your data are being analyzed by the software program.

Practical Issues in Estimating Effect Sizes

The formulas for calculating effect sizes are straightforward. In practice, however, meta-analysts face many technical issues when they

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

attempt to calculate a magnitude of effect. The most important of these is missing data, which I discussed in Chapter 4 and will return to again in the next chapter. Other issues arise because different studies use somewhat different designs and because of some unique characteristics of the effect size metrics themselves. I will describe a few of these.

Choosing a metric when studies have different designs. Some primary researchers use parametric statistics (those that assume normal distributions) and others use nonparametric statistics (ones that make no assumptions about distributions) to test and express the same relationship. For instance, this would be the case if one researcher measured intrinsic motivation in a choice study by calculating the average time each participant spent on the chosen task during a free-play period (a continuous variable dictating the use of parametric tests), and another simply recorded whether each participant did or did not choose a particular task during a free-play period (a dichotomous variable dictating use of nonparametric tests). Most often, in a research literature statistical techniques based on one set of assumptions will predominate greatly over the other. Then, the statistics from the lesser-used approach can be converted to their dominant-approach equivalents and aggregated as though they shared the dominant approach's assumptions. As long as the number of these conversions is small, there will be no great distortion of results. If there are substantive reasons to distinguish between the outcome variables or if the split between parametric and nonparametric tests is relatively even, the two sets of studies might be meta-analyzed separately.

Related to the issue of studies that use different statistical procedures is that different primary researchers sometimes convert continuous variables to dichotomous ones. For instance, some primary researchers studying the relation between individual differences and attitudes toward rape might dichotomize personality scores into high and low scoring groups. Then, they might use a *t*-test to determine if the high and low group means were different on a continuous measure of attitudes toward rape. This suggests that a *d*-index would be most appropriate to estimate the relation. However, other researchers might leave the same personality scale in its continuous form and report the correlation between them. Conveniently, the different effect size

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

metrics are easily converted from one to the other. The *r*-index can be transformed into a *d*-index using the following formula:

$$d = \frac{2r}{\sqrt{1-r^2}} \tag{10}$$

or the *d*-index into the *r*-index using

$$r = \frac{d}{\sqrt{d^2 + a}} \tag{11}$$

where

a = a correction factor to adjust for different sample sizes between the two groups.

This correction factor, *a*, can be calculated using this formula:

$$a = \frac{(n_1 + n_2)^2}{n_1 n_2} \tag{12}$$

where

all variables are defined as above.

When a chi-square statistic associated with a 2 × 2 contingency table is given, the *r*-index can be estimated as follows:

$$r = \sqrt{\frac{\chi^2}{n}} \tag{13}$$

where

χ^2 = the chi-square value associated with the comparison, and

n = the total number of observations in the comparison.

If you search the Internet using "effect size converter," you will find several websites that will allow you to easily convert between different effect size metrics.

Even though metrics can be converted easily, meta-analysts still must pick a single metric in which to describe their results. The choice

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

of how to express the effect size should be determined by which metric best fits with the measurement and design characteristics of the variables under consideration. So, *the effect size metric used should be based on the characteristics of the conceptual variables.* Therefore, an important question to ask when evaluating a research synthesis is,

> If a meta-analysis was performed, was an appropriate effect size metric used?

When we related individual differences to rape attitudes, the r-index was appropriate most often (e.g., when personality dimensions were of interest) because the two variables were conceptually continuous in nature. If a study created two artificial groups by dichotomizing the continuous individual difference measure into high and low scorers, we would calculate a d-index comparing the group means, then convert it to an r-index using Formula (11).

Estimating effect sizes when studies compare more than two groups. Suppose we find a study of interventions to promote aerobic exercise that compared three groups—say, an exercise group, an information group, and a no-intervention group. In this instance, we likely would calculate two d-indexes, one comparing exercise to no-intervention and another comparing the exercise intervention to the information intervention (we could also consider comparing the information intervention to no intervention, if this were the focus of our meta-analysis).[3] These two d-indexes are not statistically independent since both rely on the means and standard deviations of the same intervention group. However, this complicating factor is preferable to the alternative strategy of using an effect size metric associated with a multiple-group inference test. Here is why.

One effect size metric that can be used when more than two groups are being compared simultaneously involves calculating the percentage of variance in the dependent variable explained by group membership. This effect size has the initially appealing characteristic that it can be used regardless of the number of groups in the study (indeed, it can be used with two continuous measures as well). So, it is very generally applicable. However, it has the unappealing characteristic that the

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

resulting effect size tells us nothing about which of the multiple conditions has the highest mean, or, more specifically, how the values of the means are ordered and how much each differs from the others. So, identical percentages of variance explained can result from different rank ordering of, and distances between, the group means. It is then impossible for the meta-analysts to draw conclusions about how the different groups stack up relative to one another. In fact, the results might cancel one another out if we looked at single-degree-of-freedom comparisons, suggesting no differences between groups. The percentage of variance explained would not catch this. This is why it is rarely, if ever, used by meta-analysts.

Estimating effect sizes from analyses including multiple predictor variables. Another way that research design influences effect sizes involves the number of factors employed in the primary data analysis procedures. For example, a primary researcher testing the effect of homework versus no-homework on achievement might also include individual difference variables—such as the sex or previous achievement of the students, or even their pretest scores on the outcome measure—in a multi-factored analysis of variance. The primary researcher also might not report the simple means and standard deviations for the homework and no-homework groups. Meta-analysts then are faced with two choices.

First, they can calculate an effect size estimate based on the *F*-test reported by the researchers. However, this test uses an error term that has been reduced by the inclusion of the individual difference factors. This is equivalent to reducing the size of the estimate of S_{within} in the *d*-index formula. This approach creates the problem that different effect sizes going into the same quantitative synthesis are likely to be known to differ in a systematic way—that is, in how the within-group standard deviation has been calculated. Likely, if the additional factors in the analysis are associated with variance in the outcome measure (e.g., the scores on a unit test), then this study will produce a larger effect size for homework than a study that did not include these additional factors in the analysis, all else being equal.

A second approach is to attempt to retrieve the standard deviations that would have occurred had all the extraneous factors been ignored

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

(i.e., not been removed from the error term used to calculate the *F*-test). Whenever possible, this strategy should be used—that is, an attempt should be made to calculate the effect size as though the comparison of interest was the sole comparison in the analysis. The best way to do this is to contact the authors of the primary research and see if they will share the data you need. Perhaps a more realistic approach is to adjust the effect size by estimating the relationships between the additional variables and the outcome measure. Borenstein et al. (2009) present some ways to calculate these estimates. The problem here, of course, is that the resulting estimate of the effect size is only as good as the estimates of the relationships used to make adjustments.

Practically speaking, then, it is often difficult for meta-analysts to retrieve the unadjusted standard deviation estimates for the two groups if they are not given in the primary research report, nor is a simple *t*-test or one-degree-of-freedom *F*-test. In such cases, when you look for influences on study outcomes, you should either (a) leave these estimates out, if they are few, or (b) examine whether or not the number of factors included in the analysis is associated with the size of the effect. If a relation is found, you should report separately the results obtained from analyses of studies that used only the single factor of interest. So, for example, in the meta-analysis of homework research, we found one experimental study that reported the effect of homework only in an analysis of covariance with several covariates. This study's results could not be combined with studies that did not adjust for covariates. We also found other studies that presented results regarding the relation between time spent on homework and achievement only in multiple regression analyses. These could not be combined with the studies that presented simple bivariate correlations.

Adjusting for the impact of methodological artifacts. The magnitude of an effect size will also be influenced by the presence of methodological artifacts in the primary data collection procedures. Schmidt and Hunter (2015) describe 10 such artifacts that can make an effect size smaller than it might otherwise be. These include, for example, errors (lack of reliability) in the measurement of the independent and dependent variable, imperfect construct

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

validity of measures, dichotomizing of continuous variables, and restrictions in the range of sampled values.

In the case of less-reliable measures, measures with more error are less sensitive for detecting relationships involving its conceptual variables. For example, assume two personality dimensions have equal true relationships with attitudes toward rape. However, if one personality variable is measured with more error than the other, this less-reliable measure will produce a smaller correlation, all else being equal. So you might estimate the impact of the reliability of measures on effect sizes by obtaining the reliabilities (e.g., internal consistencies) of the various measures. Or, if the reliabilities of some measures were not available you could estimate the distribution (mean and standard deviation) of the reliabilities. Using procedures described by Schmidt and Hunter (2015), you could then estimate what the average effect sizes would be if all measures were perfectly reliable. You could also calculate a credibility interval, the estimated standard deviation of the disattenuated effect sizes.

Whether effect sizes should be corrected for artifacts depends first and foremost on the goal of the primary research and research synthesis. In particular, are you interested in the relationship between the constructs that underlie the measures or in what can be expected in the real world? For example, the amount of homework students do and their subsequent achievement may be imperfectly measured but if the synthesis is meant to describe what effect of homework parents, teachers, and student might expect on test scores, correcting for artifacts is inappropriate.[4] On the other hand, the meta-analysis of studies of the effect of choice on motivation might legitimately correct for unreliability in the motivation measures because they are interested in testing a theoretical notion. Error in the measurements might lead to accepting a null hypothesis when, in fact, it should be rejected.

In addition, you should keep in mind that when you correct for artifacts, your results are only as good as your estimates of the impact of the artifact. If the measures of artifacts are unreliable or you must estimate the distribution of artifact effects based on limited data, it might be good to perform a sensitivity analysis—that is, to conduct your analyses with high and low estimates of the artifact correction to see how your results differ.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Coding Effect Sizes

The statistics you need to calculate effect sizes and all the other statistics described next should be collected as part of your more general coding procedures. For example, Table 6.2 provides a simple example of the information on the statistical results of studies that might be collected by study coders. Here, the example involves experimental studies of the effects of homework on achievement. Most meta-analyses in which two conditions are being compared (having a choice among tasks, participation versus no participation in an

Table 6.2 An Example Coding Sheet for the Statistical Outcomes of Experimental Studies on the Effects of Homework on Achievement

Effect Size Estimate	
E1. What was the direction of the effect of homework on the achievement measure? + = positive – = negative	___
E2. Information about each experimental group (Note: Leave blank if not reported. *M* = Mean. *SD* = standard deviation.)	
Homework Group	
a. Pretest *M* on outcome (if any)	___ ___ ___ . ___ ___
b. Pretest *SD*	___ ___ ___ . ___ ___
c. Posttest *M* on outcome	___ ___ ___ . ___ ___
d. Posttest *SD*	___ ___ ___ . ___ ___
e. Sample size	___ ___ ___
No-Homework Group	
f. Pretest *M* on outcome (if any)	___ ___ ___ . ___ ___

(Continued)

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Table 6.2 (Continued)

Effect Size Estimate	
g. Pretest *SD*	___ ___ ___ . ___ ___
h. Posttest *M* on outcome	___ ___ ___ . ___ ___
i. Posttest *SD*	___ ___ ___ . ___ ___
j. Sample size	___ ___ ___
k. Total sample size (if not given for each group separately)	___ ___ ___ ___
E3. Information about null hypothesis significance tests	
a. Value of independent *t*-statistic (or square root of *F*-test in one-factor ANOVA)	___ ___ . ___ ___
b. Degrees of freedom for test (in the denominator)	___ ___ ___
c. *p*-value from test	< .___ ___ ___
d. Dependent *t*-statistic	___ ___ . ___ ___
e. Degrees of freedom for test (in the denominator)	___ ___ ___
f. *p*-value from test	< .___ ___ ___
g. *F*-statistic (when included in a multifactored ANOVA)	___ ___ . ___ ___
h. Degrees of freedom for denominator of *F*-test	___ ___ ___
i. *p*-value from *F*-test	< .___ ___ ___
j. # of variables in multifactored ANOVA	___
E4. Effect Size estimate	
a. What is the metric of the effect size (*d, r, OR, RR, other)*	___
b. Was an effect size calculator used to calculate this effect? 0 = No 1 = Yes	___
If yes, what calculator was used?	______________

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

exercise intervention) would look very similar. Coding sheets for correlational studies or studies relating two dichotomous variables would also be similar, but these might be even a bit simpler than my example in Table 6.2. Some of the information on the coding sheet may never be used and much of this information will be left blank. For example, when studies give the means and standard deviations, you may never use the information on the *t*-test. However, when means and/or standard deviations are missing, you will need the information on the null hypothesis significance test to calculate the *d*-index. Or if you want to examine whether the standard deviations in the experimental and control group are roughly equal, you will need this regardless of how you calculate the *d*-index. So, you might not know exactly what information is important to you until after you have begun your analysis.

COMBINING EFFECT SIZES ACROSS STUDIES

Once each effect size has been calculated, the meta-analysts next average the effects that estimate the same comparison or relationship. It is generally accepted that these averages should weight the individual effect sizes based on the number of participants in their respective samples. This is because larger samples give more precise population estimates. For example, a *d*-index or *r*-index based on 500 participants will give a more precise estimate of its underlying population effect size than will an estimate based on 50 participants. The average effect size should reflect this fact. So, while unweighted average effect sizes are sometimes presented in meta-analyses, they are typically accompanied by weighted averages.

One way to take the precision of the effect size estimate into account when calculating an average effect size is to multiply each estimate by its sample size and then divide the sum of these products by the sum of the sample sizes. However, there is a more precise procedure, first described in detail by Hedges and Olkin (1985), which has many advantages but also involves more complicated calculations.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The *d*-Index

For the *d*-index, this procedure first requires the meta-analyst to calculate a weighting factor, w_i, which is the inverse of the variance associated with each *d*-index estimate. It can be calculated taking the inverse of the result of Formula (4), or more directly by using the following formula:

$$W_i = \frac{2(n_{i1}+n_{i2})n_{i1}n_{i2}}{2(n_{i1}+n_{i2})^2 + n_{i1}n_{i2}d_i^{\,2}} \tag{14}$$

where

n_{i1} and n_{i2} = the number of data points in Group 1 and Group 2 of Study i; and

d_i = the *d*-index of the comparison under consideration.

While the formula for w_i looks imposing, it is really a simple arithmetic manipulation of three numbers available whenever a *d*-index is calculated. It also is easy to program a statistical software package to perform the necessary calculation. Programs designed to perform meta-analysis (e.g., Comprehensive Meta-Analysis, 2015) will do it for you automatically.

Table 6.3 presents the group sample sizes, *d*-indexes, and weighting factors (the w_is) associated with the results of seven hypothetical comparisons. Let us assume the seven comparisons come from experiments that compared the effects of homework versus no homework on a measure of academic achievement. All seven of the experiments produced results favoring homework assignments. The results could just as easily have come from seven comparisons of groups doing aerobic exercise or not, and the measure could be cognitive functioning. Or, the participants in one group in Table 6.3 could have been given a choice between two tasks while the other group was given no choice and the outcome could be subsequent interest in the task. It is good to look at the hypothetical data with multiple concrete examples in your head. That way you can see the conceptual similarity between the

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

examples. The key here is that you recognize that the research design in this table compares two group means on a continuous variable. If for some reason the outcome variable was a dichotomy (Did the student pass the course? Did the elderly get Alzheimer's disease? Did the subject choose the task during free time?) but the majority of outcomes were continuous, the odds or risk ratio could have been converted to a *d*-index and the study included along with the others.

To further demystify the weighting factor, note in Table 6.3 that its values equal approximately half the average sample size in a group (it becomes less similar to half the average sample size as the sample sizes in the two groups become more different). It should not be surprising, then, that the next step in obtaining a weighted average effect size involves multiplying each *d*-index by its associated w_i and dividing the

Table 6.3 An Example of *d*-Index Estimation and Tests of Homogeneity

Study	n_{i1}	n_{i2}	d_i	w_i	$d_i^2 w_i$	$d_i w_i$	Q_b Grouping
1	259	265	.02	130.98	.052	2.619	A
2	57	62	.07	29.68	.145	2.078	A
3	43	50	.24	22.95	1.322	5.509	A
4	230	228	.11	114.32	1.383	12.576	A
5	296	291	.09	146.59	1.187	13.193	B
6	129	131	.32	64.17	6.571	20.536	B
7	69	74	.17	35.58	1.028	6.048	B
5	1083	1101	1.02	544.27	11.69	62.56	

NOTE: Weighted average $d.$ = 62.56/544.27 = +.115;

$$CI_{d.95\%} = .115 \pm 1.96\sqrt{\frac{1}{544.27}} = .115 \pm .084;$$

$$Q_t = 11.69 - \frac{62.56^2}{544.27} = 4.5;$$

$Q_w = 1.69 + 2.36 = 4.05;$

$Q_b = 4.5 - 4.05 = 0.45$

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

sum of these products by the sum of the weights. This is done using the following formula:

$$\text{d.} = \frac{\sum_{i=1}^{k} d_i w_i}{\sum_{i=1}^{k} w_i} \tag{15}$$

where

k = the total number of comparisons and

all other terms are defined as above.

Table 6.3 shows the average weighted d-index for the seven comparisons is $d. = .115$.

One advantage of using the w_is as weights, rather than sample sizes, is that the w_is can also be used to generate a confidence interval around the average effect size estimate. To do this, an estimated variance for the average effect size must be calculated. First, the inverse of the sum of the w_is is found. Then, the square root of this variance is multiplied by the z-score associated with the confidence interval of interest. Thus, the formula for a 95% confidence interval is

$$\text{CI}_{\text{d.95\%}} = d. \pm z_i \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}} \tag{16}$$

where

z_i = the z-score associated with the confidence interval of interest and

all terms are defined as above.

Table 6.3 reveals that the 95% confidence interval for the seven homework comparisons encompasses values of the d-index .084 above and below the average d-index. Thus, we expect 95% of estimates of this effect to fall between $d = .031$ and $d = .199$. Note that this interval does not contain the value $d = 0$. It is this information that can be taken as a test of the null hypothesis that no relation exists in the population, in place of directly combining the significance levels of null hypothesis tests. In this

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

example, we would reject the null hypothesis that there was no difference in achievement between students who did and did not do homework.

The *r*-Index

The procedure for finding the average weighted *r*-index and its associated confidence interval is similar. Here, I will illustrate how to do this when each *r*-index is first transformed to its corresponding *z*-score, z_i. In this case, the following formula is applied:

$$z. = \frac{\sum_{i=1}^{k}(n_i - 3)z_i}{\sum_{i=1}^{k}(n_i - 3)} \qquad (17)$$

where

n_i = the total sample size for the *i*th comparison and

all other terms are defined as above.

Notice that formulas for calculating average effect sizes all follow the same form: multiply the effect size by a weight, sum the products, and divide by the sum of the weights. So, to combine the *r*-indexes directly, multiply each by its weighting factor—in this case, like the *d*-index, it is the inverse of its variance (Formula [7])—and divide the sum of this product by the sum of the weights, just as was done for the *d*-index.

To obtain a confidence interval for the average *z*-score, the formula is

$$\mathrm{CI}_{z.95\%} = z. \pm \frac{1.96}{\sqrt{\sum_{i=1}^{k}(n_\mathrm{i} - 3)}} \qquad (18)$$

where

all terms are defined as above.

To obtain a confidence interval for the *r*-indexes combined directly, simply substitute the sum of the weights in the denominator of Formula (18).

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Remember that it is important to transform your *r*-indexes to *z*-scores before you begin to combine them, especially if many of the correlations are above .25. Once the confidence interval has been established, meta-analysts convert the *z*-scores back to the correlations.

Table 6.4 presents an example of how average *r*-indexes are calculated. For example, the six correlations might come from studies relating participants' individual differences on authoritarianism and their score on a measure of rape myth acceptance. Or, the correlations might be between time spent on homework and a unit test score. Again, the key here is that both measures are continuous. The average z_i was 207 with the 95% confidence interval ranging from .195 to .219. Note that this confidence interval is quite narrow. This is because the effect size estimates are based on large samples. Note also that the *r*-to-*z* transformations result in only minor changes in two of the *r*-index values. This

Do not copy, post, or distribute

Table 6.4 An Example of *r*-Index (Transformed to *z*) Estimation and Tests of Homogeneity

Study	n_i	r_i	z_i	$n_i - 3$	$(n_i - 3)z_i$	$(n_i - 3)$ z_i^2	Q_b Grouping
1	3,505	.06	.06	3,502	210.12	12.61	A
2	3,606	.12	.12	3,603	432.36	51.88	A
3	4,157	.22	.22	4,154	913.88	201.05	A
4	1,021	.08	.08	1,018	81.44	6.52	B
5	1,955	.27	.28	1,952	546.56	153.04	B
6	12,146	.26	.27	12,143	3278.61	885.22	B
5	26,390	1.01	1.03	26,372	5462.97	1310.32	

NOTE: Weighted average *z*. = $\frac{5462.97}{26,372}$ = .207;

$$CI_{z.95\%} = .207 \pm 1.96\sqrt{26,372} = .207 \pm .012;$$

$$Q_t = 1310.32 - \frac{(5462.97)^2}{26,372} = 178.66;$$

$$Q_w = 34.95 + 50.40 = 85.35;$$

$$Q_b = 178.66 - 85.35 = 93.31.$$

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

would not be the case had the *r*-indexes been larger. As with the earlier example, $z_i = 0$ is not contained in the confidence interval. Therefore, we can reject the null hypothesis that there is no relation between participants' individual differences on authoritarianism and their scores on a measure of rape myth acceptance (or, on time spent on homework and a unit test score).

In sum, each of the effect size metrics can be averaged across studies and confidence intervals can be placed around these mean estimates. Therefore, when evaluating a research synthesis, it is important to ask,

> If a meta-analysis was performed, (a) were average effect sizes and confidence intervals reported and (b) was an appropriate model used to estimate the independent effects and the error in effect sizes?

A Note on Combining Slopes From Multiple Regressions

Up to this point, the procedures for combining and comparing study results have assumed that the measure of effect is a difference between means, a correlation, or an odds ratio. However, regression analysis is a commonly used technique in the social sciences, particularly in nonexperimental studies where many variables are used to predict a single criterion. Similar to the standardized mean difference or correlation coefficient, the regression coefficient, *b,* or the standardized regression coefficient, β, are also measures of effect size. β will typically be of most interest to meta-analysts because, like the *d*-index and *r*-index, it standardizes effect size estimates when different measures of the same conceptual variable are used in different studies. β represents the change in a standardized predictor variable, controlling for all other predictors, given one standard unit change in the criterion variable.

Meta-analyses using regression coefficients as effect sizes are difficult to conduct for a variety of reasons. First, with regard to using the unstandardized *b*-weight, this is like using raw score differences as measures of effect—the scales of the predictor and outcome of interest

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

typically vary across studies. Directly combining them can lead to uninterpretable results. This problem can be overcome by using β, the fully standardized estimate of the slope for a particular predictor.[5] But still, the other variables included in models using multiple regression generally differ from study to study (note the related earlier discussion about multifactored analyses of variance). Each study may include different predictors in the regression model and, therefore, the slope for the predictor of interest will represent a different partial relationship in each study (Becker & Wu, 2007). For example, in our meta-analysis of homework and achievement, we found numerous studies that performed analyses of the relationship between time spent on homework and achievement that reported β. However, each was based on a regression model that included different additional variables. This made it questionable that the βs should be directly combined. So, rather than average them, we described these studies' individual βs and the range of β-values across the studies. These were overwhelmingly positive, were generally based on very large samples, and used a variety of achievement outcome measures. As such, they strengthened our claim about the positive effects of homework on achievement that was based on the few small studies that purposively manipulated homework and tested its effect on a single limited outcome measure, unit test scores.

Regression slopes can be directly combined when (a) the outcome and predictor of interest are measured in a similar fashion across studies, (b) the other predictors in the model are the same across studies, and (c) the predictor and outcome scores are similarly distributed (Becker, 2005). It is rare that all three of these assumptions are met; typically, measures differ across studies and regression models are diverse in terms of which additional variables are included in them.

The Synthesis Examples

Both the standardized mean difference and the correlation coefficient measures of effect size were used in the synthesis examples. In the synthesis of the effects of homework, the *d*-index was used to express the findings from comparisons that purposively manipulated homework

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

and then measured the difference in terms of unit test scores. The weighted average *d*-index across five studies was *d*. = .60, with a 95% confidence interval encompassing values from *d* = .38 to *d* = .82. Clearly, then, the null hypothesis could be rejected. The homework research synthesis also used correlation coefficients to estimate the relationship between student or parent reports of the amount of time spent on homework and a variety of measures of achievement. Of 69 such correlations, 50 were positive and 19 were negative. The weighted average correlation was *r* = .24 with a very narrow 95% confidence interval, encompassing the values between .24 and .25. The confidence interval was so small because of the large number of participants in these studies; the adjusted mean sample size in the studies was 7,742.

The meta-analysis of individual differences and attitudes toward rape also used correlation coefficients as the measure of the strength of relationships. Among the many correlations involving individual differences, we found, for example, that across 15 correlations older participants were more accepting of rape than younger ones, average *r*. = .12 (95% CI = .10–.14).

The meta-analyses on (a) interventions to increase aerobic exercise among adults and (b) the effects of choice on intrinsic motivation used the standardized mean difference to measure effects. The weighted average *g*-index across 29 studies indicated that adults who participated in the interventions revealed improvements in attention and processing speed, *g*. = .158 (95% CI = .055–.260), executive functioning, *g*. = .123 (95% CI = .021–.225), and memory, *g*. = .128 (95% CI = .015–.241). The average weighted effect size for the 47 estimates of the impact of choice on measures of intrinsic motivation was *d*. = .30 (95% CI = .25–.35), indicating choice led to greater intrinsic motivation.

ANALYZING VARIANCE IN EFFECT SIZES ACROSS FINDINGS

The analytic procedures described thus far have illustrated how to estimate effect sizes, average them, and use the confidence interval surrounding the average to test the null hypothesis that the difference

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

between two means or the size of a correlation is 0. Another set of statistical techniques helps meta-analysts discover why effect sizes vary from one comparison to another. In these analyses, the effect sizes found in the separate comparisons are the *dependent* or predicted variables and the characteristics of the comparisons are the predictor variables. The meta-analysts ask whether the magnitude of relation between two variables in a comparison is affected by the way the study was designed or carried out.

One obvious feature of the effect sizes in Tables 6.3 and 6.4 is that they vary from comparison to comparison. An explanation for this variability is not only important, but also represents the most unique contribution of research synthesis. By performing an analysis of differences in effect sizes, the meta-analyst can gain insight into the factors that affect the strengths of relationships even though these factors may have never been studied in a single experiment. For instance, assume that the comparisons were looking at the effects of homework and the first four studies listed in Table 6.3 were conducted in elementary schools while the last three studies were conducted in high schools. Is the effect of homework different for students at different grades? This question could be addressed through the use of the analytic techniques described next, even though no single study included both elementary and high school students and tested to see if the grade level of students moderated the effect of homework.

The techniques that follow are a few examples of many procedures for analyzing variance in effect sizes. I do not cover some of the more complex synthesis techniques but will return to them after exposition of the most frequently used meta-analysis techniques.

Traditional Inferential Statistics

One way to analyze the variance in effect sizes is to apply the traditional inference procedures that are used by primary researchers. Meta-analysts interested in whether an exercise intervention's effects on older adults' cognitive functioning were stronger for males than for females might do a *t*-test on the difference between effect sizes found in comparisons exclusively using males versus comparisons exclusively using females. Or, if the meta-analysts were interested in whether the

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

intervention effect size was influenced by the length of the intervention and the measurement of cognitive functioning, the meta-analysts might correlate the length of treatment in each comparison with its effect size. In this instance, the predictor and dependent variables are continuous, so the significance test associated with the correlation coefficient would be the appropriate inferential statistic. For more complex questions, a synthesist might categorize effect sizes into multifactor groupings—for instance, according to the gender and age of participants—and perform an analysis of variance or multiple regression on effect sizes. For Table 6.3, if a one-way analysis of variance were conducted comparing the first four *d*-indexes with the last three *d*-indexes, the result would not be statistically significant.

Standard inference procedures were the techniques initially used by some meta-analysts for examining variance in effects. Glass et al. (1981) detailed how this approach is carried out. However, at least two problems arise with the use of traditional inference procedures in meta-analysis. The first is that traditional inference procedures do not test the hypothesis that the variability in effect sizes is due solely to sampling error (recall the discussion earlier in this chapter). Therefore, the traditional inference procedures can reveal associations between design characteristics and effect sizes without determining first whether the overall variance in effects is greater than that expected by sampling error alone.

Also, because effect sizes can be based on different numbers of data points (sample sizes), they can have different sampling variances associated with them—that is, they are measured with different amounts of error, or differing levels of precision. If this is the case (and it often is), then the effect sizes violate the assumption of homogeneity of variance that underlies traditional inference tests. For these two reasons, traditional inferential statistics are no longer used when performing a meta-analysis.

Comparing Observed to Expected Variance: Fixed-Effect Models

In place of traditional procedures, several approaches have gained acceptance. One approach is called the *fixed-effect model*. I will

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

explain this simplest model first and then explain a second more complex model, called the *random-effects model.* The fixed-effect model compares the variation in the observed effect sizes with the variation expected if only error due to the sampling of participants were causing differences in effect size estimates. In other words, it makes the assumption that there is one value of the effect size underlying all the observations and the only thing making the observations different is differences in the participants sampled into each study. This approach involves calculating (a) the observed variance in the effect sizes from the known findings and (b) the expected variance in these effect sizes given that all are estimating the same underlying population value. Sampling theory allows us to calculate precise estimates of how much sampling variation to expect in a group of effect sizes if only differences between the participants is making the effect sizes different. This expected value is a function of the average effect size estimate, the number of estimates, and their sample sizes.

The meta-analysts then compare the observed with the expected variance. If the variance estimates are deemed not to differ then sampling error of participants is the simplest explanation for the variance in effect sizes. If they are deemed different—that is, if the observed variance is (significantly) greater than that expected due to sampling error of participants, then the meta-analysts begin the search for systematic influences on effect sizes. This is done by grouping the effect sizes and asking whether the group averages are more different than sampling error alone would predict.

Homogeneity Analyses

A homogeneity analysis is a formal way to compare the observed variance to that expected from sampling error. It involves the calculation of how probable it is that the variance exhibited by the effect sizes would be observed if only sampling error was making them different. This is the approach used most often by meta-analysts, so I will provide a few more of its details.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Homogeneity analysis first asks the question, "Is the observed variance in effect sizes statistically significantly different from that expected by sampling error alone?" If the answer is "no," then some statisticians advise that the meta-analysts stop the analysis there. After all, chance or sampling error is the simplest and most parsimonious explanation for why the effect sizes differ. If the answer is yes—that is, if the effect sizes display significantly greater variability than expected by chance, the meta-analysts then begin to examine whether study characteristics are systematically associated with variance in effect sizes. Some meta-analysts believe that the search for moderators should proceed regardless of whether sampling error is rejected as a plausible sole cause of variability in effect sizes *if* there are good theoretical or practical reasons for choosing moderators. This is the approach I usually take. Regardless of the approach you prefer, when evaluating a research synthesis, it is important to ask,

If a meta-analysis was performed, was the homogeneity of effect sizes tested?

Suppose a meta-analysis reveals a homogeneity statistic that has an associated *p*-value of .05. This means that only 5 times in 100 would sampling error create this amount of variance in effect sizes. Thus, the meta-analysts would reject the null hypothesis that sampling error alone explains the variance in effect sizes and they would begin the search for additional influences. They would then test whether study characteristics explain variation in effect sizes. Studies would be grouped by common features, and the average effect sizes for groups would be tested for homogeneity in the same way as the overall average effect size.

An approach to homogeneity analysis will be described that was introduced simultaneously by Rosenthal and Rubin (1982) and Hedges (1982). The formula presented by Hedges and Olkin (1985; also see Hedges, 1994) will be given here and the procedures using *d*-indexes will be described first.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

The d-*index.* In order to test whether a set of *d*-indexes is homogeneous, the meta-analysts must calculate a statistic Hedges and Olkin (1985) called Q_t. The formula is as follows:

$$Q_t = \sum_{i=1}^{k} w_i d_i^{\,2} - \frac{\left(\sum_{i=1}^{k} w_i d_i\right)^2}{\sum_{i=1}^{k} w_\mathrm{i}} \tag{19}$$

where all terms are defined as above.

The *Q*-statistic has a chi-square distribution with $k - 1$ degrees of freedom, or, one less than the number of comparisons. The meta-analysts refer the obtained value of the total *Q* statistic, Q_t, to a table of (upper tail) chi-square values. If the obtained value is greater than the critical value for the upper tail of a chi-square at the chosen level of significance, the meta-analysts reject the hypothesis that the variance in effect sizes was produced by sampling error alone. Table 6.5 presents the critical values of chi-square for selected probability levels.

Table 6.5 Critical Values of Chi-Square for Given Probability Levels

	Upper Tail Probabilities					
DF	**.500**	**.250**	**.100**	**.050**	**.025**	**.010**
1	.455	1.32	2.71	3.84	5.02	6.63
2	1.39	2.77	4.61	5.99	7.38	9.21
3	2.37	4.11	6.25	7.81	9.35	11.3
4	3.36	5.39	7.78	9.49	11.1	13.3
5	4.35	6.63	9.24	11.1	12.8	15.1
6	5.35	7.84	10.6	12.6	14.4	16.8
7	6.35	9.04	12.0	14.1	16.0	18.5
8	7.34	10.2	13.4	15.5	17.5	20.1
9	8.34	11.4	14.7	16.9	19.0	21.7

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

	Upper Tail Probabilities					
DF	**.500**	**.250**	**.100**	**.050**	**.025**	**.010**
10	9.34	12.5	16.0	18.3	20.5	23.2
11	10.3	13.7	17.3	19.7	21.9	24.7
12	11.3	14.8	18.5	21.0	23.3	26.2
13	12.3	16.0	19.8	22.4	24.7	27.7
14	13.3	17.1	21.1	23.7	26.1	29.1
15	14.3	18.2	22.3	25.0	27.5	30.6
16	15.3	19.4	23.5	26.3	28.8	32.0
17	16.3	20.5	24.8	27.6	30.2	33.4
18	17.3	21.6	26.0	28.9	31.5	34.8
19	18.3	22.7	27.2	30.1	32.9	36.2
20	19.3	23.8	28.4	31.4	34.2	37.6
21	20.3	24.9	29.6	32.7	35.5	33.9
22	21.3	26.0	30.8	33.9	36.8	40.3
23	22.3	27.1	32.0	35.2	38.1	41.6
24	23.3	28.2	33.2	36.4	39.4	43.0
25	24.3	29.3	34.4	37.7	40.6	44.3
26	25.3	30.4	35.6	38.9	41.9	45.6
27	26.3	31.5	36.7	40.1	43.2	47.0
28	27.3	32.6	37.9	41.3	44.5	48.3
29	28.3	33.7	39.1	42.6	45.7	49.6
30	29.3	34.8	40.3	43.8	47.0	50.9
40	49.3	45.6	51.8	55.8	59.3	63.7
60	59.3	67.0	74.4	79.1	83.3	88.4
	.500	**.750**	**.900**	**.950**	**.975**	**.990**
	Lower Tail Probabilities					

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

For the set of comparisons given in Table 6.3, the value of Q_t equals 4.5. The critical value for chi-square at $p < .05$ based on 6 degrees of freedom is 12.6. Therefore, the hypothesis that sampling error explains the differences in these *d*-indexes cannot be rejected.

The procedure to test whether a methodological or conceptual distinction between studies explains variance in effect sizes involves three steps. First, a *Q*-statistic is calculated separately for each subgroup of comparisons. For instance, to compare the first four *d*-indexes in Table 6.3 with the last three, a separate *Q*-statistic is calculated for each grouping. Then, the values of these *Q*-statistics are summed to form a value called Q_w, or *Q*-within. This value is then subtracted from Q_t to obtain the *Q* statistic for the difference between the two group means, Q_b, or Q-between:

$$Q_b = Q_t - Q_w \tag{20}$$

where

all terms are defined as above.

The statistic Q_b is used to test whether the *average* effects from the two groupings are homogenous. It is compared to a table of chi-square values using as degrees of freedom one less than the number of groupings. If the average *d*-indexes are homogeneous, then the grouping factor does not explain variance in effects beyond that associated with sampling error. If Q_b exceeds the critical value, then the grouping factor is a significant contributor to variance in effect sizes.

In Table 6.3 the Q_b comparing the first four and last three *d*-indexes is .45. This result is not significant with one degree of freedom. So, if the first four effect sizes were taken from studies of the effect of homework on achievement using elementary school students and the last three using high school students, we could not reject the null hypothesis that effect sizes were equal in the two populations of students.

The r*-index.* The analogous procedure for performing a homogeneity analysis on *r*-indexes transformed to *z*-scores involves the following formula:

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

$$Q_t = \sum_{i=1}^{k}(n_i - 3)z_i^2 - \frac{\left[\sum_{i=1}^{k}(n_i - 3)z_i\right]^2}{\sum_{i=1}^{k}(n_i - 3)} \qquad (21)$$

where

all terms are defined as above.

To compare groups of *r*-indexes, Formula (21) is applied to each grouping separately, and the sum of these results, Q_w, is subtracted from Q_t to obtain Q_b.

The results of a homogeneity analysis using the *z*-transforms of the *r*-indexes are presented in Table 6.4. The Q_t value of 178.66 is highly significant, based on a chi-square test with 5 degrees of freedom (the number of correlation minus one). While it seems that a range of *r*-indexes from .06 to .27 is not terribly large, Q_t tells us that, given the sizes of the samples on which these estimates are based, the variation in effect sizes is too great to be explained by sampling error alone. Something other than sampling of participants likely is contributing to the variance in *r*-indexes.

Suppose we know that the first three correlations in Table 6.4 are from samples of high school students and the last three are from elementary school students. A homogeneity analysis testing the effect of grade level on the magnitude of *r*-indexes reveals a Q_b of 93.31. This value is highly statistically significant, based on a chi-square test with one degree of freedom. For high school students the average weighted *r*-index is .253, whereas for elementary school students it is $r = .136$. Thus, the null hypothesis can be rejected and the grade level of the student is one potential explanation for the variation in *r*-indexes.

Comparing Observed and Expected Variance: Random-Effects Models

An important decision you will make when conducting a meta-analysis involves whether a fixed-effect or random-effects model should be used to calculate the variability in effect size estimates averaged

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

across studies. As I discussed above, fixed-effect models calculate only error that reflects variation in studies' outcomes due to the sampling of participants. However, other features of studies also can be viewed as influences on outcomes. For example, the studies in a synthesis of homework may vary by the length of the assignment and/or subject matter. Exercise interventions may vary in their intensity or modality. Choices may vary in number or domain. These variations will cause variation in effect sizes not due to sampling of participants. However, they are not error in the sense of being chance because even though they may at first be unexplained they may also be systematic in ways we are not aware of. For example, more-intense exercise interventions may improve cognitive functioning more than less-intense interventions.

For this reason, in many cases it may be most appropriate to treat studies as randomly sampled from a population of all studies. The variation that might be added to the estimate of error due to variations in study methods is ignored when a fixed-effect model is used. In a random-effects model (Raudenbush, 2009), study-level variance is assumed to be present as an additional source of random influence. The question you must answer, then, is whether you believe the effect sizes in your data set are noticeably affected by study-level influences.

Regrettably, there are no hard-and-fast rules for making this determination. Overton (1998) found that in the search for moderators, fixed-effect models may seriously underestimate error variance and random-effects models may seriously overestimate error variance when their assumptions are violated. Thus, neither can be chosen because it is statistically more justified. In practice, many meta-analysts opt for the fixed-effect assumption because it is analytically easier to manage. But some meta-analysts argue that fixed-effect models are used too often when random-effects models are more realistic, such as when interventions like homework or exercise programs can be expected to have different empirical realizations from one study to another in ways that will influence their effectiveness. Others counter this argument by claiming that a fixed-effect model can be applied if a thorough, appropriate search for moderators of effect sizes is part of the analytic strategy—that is, if the meta-analysts examine the systematic effects of study-level

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

influences—and in this way make moot the issue of random effects at the study level. The fixed-effect model may also be favored if the number of effect sizes is small, making it difficult to achieve a good estimate of variation in the effect sizes at the study level.

What should your decision be based on? One approach is to decide based on the outcome of the test of homogeneity of effects using a fixed-effect model; if the hypothesis of homogeneous effects is rejected under the fixed-effect assumption, then you switch to a random-effects model. However, as Borenstein et al. (2009) argue, this strategy is discouraged; it is based on statistical outcomes, not on the conceptual characteristics of your studies. Many researchers interested in evaluating applied interventions (such as homework) often choose the random-effects model because they believe that random sampling of studies is more descriptive of their real-world circumstances and also will lead to a more conservative conclusion about the range of impacts the intervention might have (because the estimate of the variation around the average estimate is larger using the random-effects model). So, if you suspect a large influence of study-level sources of random error, then a random-effects model is most appropriate in order to take these sources of variance into account.

Other researchers studying basic social processes—processes that likely do not change greatly due to the contexts in which they are being studied (such as, perhaps, tests of reaction times)—tend to favor fixed-effect models. Hedges and Vevea (1998) stated that fixed-effect models are most appropriate when the goal of the research is "to make inferences only about the effect size parameters in the set of studies that are observed (or a set of studies identical to the observed studies except for uncertainty associated with the sampling of subjects)" (p. 3). In studies of basic processes, this type of inference might suffice, because you make the extra-statistical assumption that the relationship you are studying is largely insensitive to its context.

To summarize then, you might consider applying the following rules:

- Do not use the outcome of a fixed-effect homogeneity analysis to decide whether a random-effects analysis is called for. The decision should be based on the nature of the research question.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

- In most instances where interventions are being evaluated or the research takes place in real-world contexts that vary from one another in important ways, random-effects models should be favored. However, if the number of studies being combined is small, consider using a fixed-effect model; the estimate of study-level variance will be too rough.
- In most instances where laboratory studies of basic processes are being combined, fixed-effect models should be appropriate. Here, the context of the study (study-level variations) should be less consequential to study findings.

Which model of effects you use and the set of assumptions your choice is based on needs to be incorporated into the interpretation and discussion of your findings. I will return to the issue of interpreting fixed-effect and random-effects models in Chapter 7.

Calculating random-effects estimates of the mean effect size, confidence intervals, homogeneity statistics, and moderator analyses is computationally complex. Because of this complexity, the formulas I have provided in this chapter are for fixed-effect models. I will not go into the calculation of the variance estimate in random-effects models (see Borenstein et al., 2009, if you are interested) but conceptually it involves calculating the variation in effect sizes (using the effect size as the unit of analysis) and adding this to the variation due to sampling of participants (the fixed-effect). Thankfully, the statistical packages developed specifically for meta-analysis and the program macros associated with more general statistical packages allow you to conduct analyses using both fixed-effect and random-effects assumptions.

I^2: The Study-Level Measure of Effect

It may have occurred to you that meta-analysts point out the shortcomings of null hypothesis significance testing but then use it to test whether groups of studies have significantly different average effect sizes. This is only partially true. Certainly, a good meta-analysis presents the confidence interval around overall estimates of effect and for all subgroups when a moderator of effects is tested. A measure of effect

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

also exists for quantifying the percentage of the variance in a set of studies that is due to the studies themselves and not sampling error. This statistic is called I^2 and is calculated as follows:

$$I^2 = \sum \frac{Q - df}{Q} \times 100\% \tag{22}$$

where

all quantities are defined as above.

I^2 tells you what portion of the total variance in the effect sizes is due to variance between the studies. The Cochrane Collaboration (Deeks, Higgins, & Altman, 2008) gives a rough guide to when the percentage of study variance may be important. In addition to the significance of the Q-statistic, it suggests that I^2 below 40% might not be important while I^2 above 75% suggests considerable heterogeneity.

Statistical Power in Meta-Analysis

The above discussion leads naturally into a consideration of the power of meta-analyses to detect effects. Meta-analyses have different statistical power for answering its multiple questions. First, meta-analysts ask the question, "What is the average effect size and the precision of this estimate, or, alternatively, with what certainty can we reject the null hypothesis?" The answer to this question will depend on the model, fixed or random, used to estimate the expected variation in effects. When a fixed-effect model is used, we can say with certainty that the power of the meta-analysis to detect an effect and the precision of the estimate will be greater in the meta-analysis than in any one or any subset of the primary studies going into the research synthesis. This is because the meta-analytic estimate will always be based on a larger sample of participants. If the assumption of the fixed-effect model is true (i.e., sampling error alone is making sample estimates different) the meta-analysis estimate will always be more precise.

However, this is not necessarily true when a random-effects model is employed. Here, the variability due to variations in study

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

characteristics must be added to sampling error at the participant level. This source of variance is not present in any one study. So, if study-level variance is large it is possible when we calculate the precision of the average effect size that the precision of the individual studies (or one or some of them) can be greater than the precision of the meta-analytic effect size estimate. You can think of it this way: if the estimate of study-level variation adds nothing to the participant-level variance, then a fixed-effect and random-effects model will provide the same estimates of variability (equal to participant sampling alone) and meta-analytic estimates of effect will always be more precise than any single-study estimate. As the study-level variability moves away from a zero contribution, the precision of the meta-analytic estimate decreases and at some point, depending on the amount of study-level variability and the number and sample size of the primary studies, may become less precise than any single study estimate.

Next, meta-analysts ask whether there is sufficient power to detect a significant *Q*-statistic, or to reject the null hypothesis that sampling of participants alone is making the effect sizes different. Similar to power analysis with primary data, the power to detect a difference between an observed *Q*-statistic and an expected one is a function of the number of effect sizes you have, the sample sizes contributing to those effects sizes, the size of the expected study-level variation in effects (the I^2) as well as how well the effects conform to the necessary statistical assumptions (e.g., normal distribution).

Finally, meta-analysts might be interested in the power to detect differences between groups of studies: "Was the average effect in the group of Studies A different from the average effect in the group of Studies B?" This power analysis requires a variation on the analyses described in the last paragraph.

Conducting power analysis in meta-analysis often has a different purpose from that in primary research. After all, meta-analysts do not do power analysis to help decide how many studies to run. Perhaps, if an existing literature contains a very large number of studies, the meta-analyst might conduct an a priori power analysis to determine how many studies to sample from it. Otherwise meta-analytic power analyses are most informative as guides to interpretation. The power of meta-analytic tests can be very low, especially for tests of moderators

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

of study effects when a random-effects model is used and the number of studies is small. By conducting such an analysis, the interpretation of the results can include the possibility that accepting the null hypothesis might lead to a Type II error.

Meta-Regression: Considering Multiple Moderators Simultaneously or Sequentially

Homogeneity statistics can become unreliable and difficult to interpret when the meta-analysts wish to test more than one moderator of effect sizes at a time. Hedges and Olkin (1985) present one technique for testing multiple moderators. The model uses simultaneous or sequential tests for homogeneity. It removes the variance in effect sizes due to one moderator and then removes from the remaining variance any additional variance due to the next moderator. So, for example, if we were interested in whether the sex of the student influenced the effect of homework on achievement after controlling for the student's grade level, we would first test grade level as a moderator, then test the student's sex as a moderator *within* each grade-level category.

This procedure can be difficult to apply because characteristics of studies are often correlated with one another and the number of effect sizes in categories of interest rapidly becomes small. For example, suppose we wanted to test whether the effect of homework on achievement is influenced by both the grade level of students and the type of achievement measure. We might find that these two study characteristics are often confounded—more studies of high school students used standardized tests while more studies of elementary school students used class grades. Studies of homework with elementary school students using standardized tests may be rare. The problem would get even worse if yet a third variable were added to the mix.

Another statistical approach to testing multiple moderators of effect sizes simultaneously or sequentially is called *meta-regression*. As the name implies, this approach is the meta-analysis analog to multiple regression. In meta-regression, the effect sizes are the criterion variables and the study characteristics are the predictors (Hartung, Knapp, & Sinha, 2008). Meta-regression shares with multiple regression all the

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

problems regarding the interpretation of the analysis' output when the predictors are intercorrelated (a likely characteristic of research synthesis data) and when the number of data points (effect sizes in meta-regression) are small.

Still, meta-regression is becoming more popular, especially now that meta-analysis programs are available to help you do them. One important consideration regarding when to use meta-regression involves the effect sizes that serve as the dependent variables. Remember that the regression analysis makes the assumption that the effect sizes are independent of one another. In Chapter 4 I discussed the units of analysis in research synthesis and some strategies for minimizing multiple outcomes that come from the same sample of participants. In meta-regression it is not unusual for the outcome rather than the sample to be used as the independent unit. This requires adjustments lest the estimates of error appear to be more precise than they actually are (see Hedges, Tipton, & Johnson, 2010).

Another approach to addressing the intercorrelation of study characteristics is to first generate homogeneity statistics for each characteristic separately, by repeating the calculation of *Q*-statistics. Then, when the results concerning moderators of effect sizes are interpreted, the meta-analysts also examine a matrix of intercorrelations among the moderators. This way, the meta-analyst can alert readers to study characteristics that may be confounded and draw inferences with these relations in mind. For example, we followed this procedure in the meta-analysis of the effects of choice on intrinsic motivation. We found that the effect of giving choices influenced children's intrinsic motivation more positively than adults' motivation. But we found also that the age of the participant was associated with the setting in which the choice experiment was conducted; studies with adults were more likely to be conducted in a traditional lab setting than were studies with children. This means that the different effect of choice on motivation for children and adults might not be due to the participants' age, but rather to where the study was conducted.

In sum, then, you need to make many practical decisions when conducting a meta-analysis, and the guidelines for making these are not as clear as we would like. While it is clear that a formal analysis of the variance in effect sizes is an essential part of any research synthesis

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

containing large numbers of comparisons, it is also clear that you must take great care in the application of these statistics and in the description of how they were applied.

Using Computer Statistical Packages

Needless to say, calculating average weighted effect sizes and homogeneity statistics by hand is time-consuming and prone to error. Today, it is unheard of for meta-analysts to compute statistics for themselves, as I have done in the previous examples. Still, it is good for you to examine my examples carefully and conduct the calculation yourself, so that you understand them. Then, the output of computer packages should be more interpretable by you and you should be more able to notice any errors that might have occurred.

Conveniently, the major computer statistics packages have macros developed that allow their use to conduct meta-analysis. For example, meta-analyses can be run using Excel spreadsheets (Neyeloff, Fuchs, & Moreira, 2012). David Wilson's very helpful website provides free macros for use with the SPSS, STATA, and SAS software packages. These packages are generally familiar to most social scientists. A book is available that shows how to use the statistical package R to conduct meta-analysis (Chen & Peace, 2013; see also http://cran.r-project.org/web/views/MetaAnalysis.html for a useful compendium of R programs). A free (though support comes at a cost) program dedicated to meta-analysis alone is called RevMan (http://tech.cochrane.org/revman/download). There are also stand-alone meta-analysis packages that can be purchased such as Comprehensive Meta-Analysis (2015) that will produce all the results for you, and give you many options for how to carry out your analyses.

Regardless of how the statistics are calculated, when evaluating a research synthesis, you should ask,

> Were (a) study design and implementation features along with (b) other critical features of studies, including historical, theoretical, and practical variables, tested as potential moderators of study outcomes?

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

SOME ADVANCED TECHNIQUES IN META-ANALYSIS

Several more advanced approaches to meta-analysis have emerged in recent years. These typically require more advanced statistical knowledge and complex calculations than can be covered in an introductory textbook. Below, I will provide a brief conceptual introduction to some of the approaches receiving the most attention. Because the complex meta-analysis techniques require full treatment to be applied and are still used relatively infrequently I will not dwell on them here. If you are interested in more advanced techniques, you should first examine these in more detailed treatments, especially those given in Cooper et al. (2009) and the references provided below.

Hierarchical Linear Modeling

One new approach to meta-analysis involves using hierarchical linear modeling (Raudenbush & Bryk, 2001). This approach treats study outcomes as nested data; for example, students' achievement scores can be viewed as influenced by (nested within) classroom-level variables that are themselves influenced by school characteristics and at a higher level still by the community the school is in. In the case of meta-analysis, a study outcome (an individual effect size) can be viewed as nested within a sample of participants who in turn are nested within a study (and even within a laboratory that has conducted multiple studies). Again, the computations for the analyses are complex but this approach is conceptually appealing and meta-analyses using the hierarchical linear modeling approach are used increasingly frequently.

Model-Based Meta-Analysis

The statistical procedures for meta-analysis described so far apply to synthesizing two-variable relationships from experimental and descriptive research. Meta-analysis methodologists are working to extend statistical synthesis procedures to more-complex ways to express the

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

relations between variables. Previously, I discussed the difficulties in synthesizing the effect sizes associated with a variable that was included in a multiple regression. But what if the question of interest involves integrating the output of *entire* regression equations? For example, suppose we were interested in how five personality variables (perhaps those in the five-factor model) jointly predicted attitudes toward rape? Here, we would want to develop from a meta-analysis a regression equation, or perhaps a structural equation model, based on the results of a set of studies. To do so, we would need to integrate results of studies concerning not one correlation between the variables but rather an entire matrix of correlations relating all the variables in the model of interest to us. It is this correlation matrix that forms the basis of the multiple regression model.

The techniques used to do this are still being explored, as are the problems meta-analysts face in using them. For example, can we simply conduct separate meta-analyses for each correlation coefficient in the matrix and then use the resulting matrix to generate the regression equation? The answer is "probably not." The individual correlations would then be based on different samples of participants and a regression analysis using them can produce nonsensical results, such as prediction equations that explain more than 100% of the variance in the criterion variable. Still, there are circumstances under which these applications of meta-analysis to complex questions can produce highly informative results. Becker (2009) presents an in-depth examination of the promise and problems involved in model-driven meta-analysis.[6]

Bayesian Meta-Analysis

Another approach to meta-analysis involves applying Bayesian statistics rather than the frequentist approach used in the statistics described in this book. In a Bayesian approach (Sutton & Abrams, 2013; Sutton et al., 2000), the researcher must first establish a prior estimation of the parameters of the effect size. These can include both the magnitude and the distribution of effect sizes. The Bayesian priors can be based on past research, and not necessarily on research that used identical conceptual variables or empirical realizations. For example, the prior estimation of

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

the effect of an exercise intervention might be based on other interventions to improve cognitive functioning, such as puzzle solving. Or, the estimation might be based on samples drawn from other populations (e.g., using adult samples to estimate the effect of choice on children's motivation) or even on subjective beliefs and personal experience (e.g., teachers' thoughts on the degree to which homework affects achievement). The meta-analysis then tells the synthesists how these prior beliefs should change in light of the new empirical evidence. The need for prior estimations in Bayesian analyses is seen as both a strength and a weakness of the approach. The computations for Bayesian analyses are also very complex and less intuitively accessible than the traditional meta-analysis methods but can yield trustworthy and interpretable results (Jonas et al., 2013).

Meta-Analysis Using Individual Participant Data

The most desirable technique for combining results of independent studies is to have available and to integrate the raw data from each relevant comparison or estimate of a relationship (Cooper & Patall, 2009). Then, the individual participant data (IPD) can be placed into a new primary data analysis that employs the comparison that generated the data as a blocking variable. When IPD are available, the meta-analysis can perform subgroup analyses that were not conducted by the initial data collectors in order to

- Check data in the original studies,
- Ensure that the original analyses were conducted properly,
- Add new information to the data sets,
- Test with greater power variables that moderate effect sizes, and
- Test for both between-study and within-study moderators.

Obviously, instances in which the integration of IPD can be achieved are rare. IPD are seldom included in research reports, and attempts to obtain raw data from researchers often end in failure. However, the incentives and requirements for sharing data are increasing, as conditions both for receiving research support and for

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

publishing findings. If IPD are retrievable, the meta-analyst still must overcome the use of different metrics in different studies, an important limit to the ability to statistically combine the results. Also, meta-analyses using IPD can be expensive because of the recoding involved in getting the data sets into similar form and content. So it is unlikely that meta-analyses using IPD will be replacing the meta-analysis techniques described previously any time soon. Still, meta-analysis of IPD is an attractive alternative, one that has received considerable attention in the medical literature, and likely will become more attractive as the availability of raw data sets improves. Also, methods are appearing that allow synthesists to use both IPD from some studies and aggregate data from others (Pigott, 2012).

CUMULATING RESULTS ACROSS META-ANALYSES

The terms *cumulative* or *prospective meta-analyses* are used to refer to meta-analyses that are updated as new evidence on a topic becomes available. The methods for conducting the new analyses can be the same as those used originally, or can be changed, perhaps to reflect advances in meta-analytic methods or to conduct new analyses that time and experience suggest are warranted; for example, looking at a new moderators variable that recent theorizing suggests might influence results. Many cumulative meta-analyses include the year of the study as a moderating variable to determine whether the evidence suggests the impact of the treatment or intervention is changing over time. Cumulative meta-analyses are much more frequently encountered in the medical than social sciences. In fact, the Cochrane Collaboration (2015) requires that synthesists who submit to its database commit to updating the reports as new information appears.

Overviews of reviews. Overviews of reviews, sometimes also called *reviews of reviews, umbrella reviews,* or *meta-reviews,* compile evidence from multiple research syntheses. Cooper and Koenka (2012) catalogued several reasons why an overview of reviews might be undertaken. These included (a) to summarize evidence from more than one research synthesis focused on the same or overlapping

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

research problems or hypotheses, (b) to compare findings and resolve discrepancies in the conclusions drawn in more than one research synthesis, and (c) to catalog the mediators and moderators tested in research syntheses on the same research problem. Like all research syntheses, there are sound methods for conducting an overview of reviews that are unique to them. For example, overviewers must evaluate the quality of the constituent research syntheses.

Overviews have their limitations as well. For example, the studies included in an overview of reviews can be quite old, considering not only that the studies must be conducted, but also that the review of studies then must be conducted and this is the evidence in the overview. Still, the same forces that are giving rise to the need for research syntheses, the expanding research literature, will also provide impetus for a growing appearance of overviews of reviews.

Second-order meta-analysis. One type of overview is called a *second-order meta-analyses.* It involves using the outcomes of meta-analyses as the data in yet another meta-analysis (Schmidt & Hunter, 2015). In second-order meta-analyses the average effects found in meta-analyses conducted in the same problem area are themselves combined. Obviously, second-order meta-analysis is used when neither the IPD nor even the study-level results from the constituent meta-analyses can be retrieved.

One problem faced by second-order meta-analysts (as well as any overviewer, only more formally) is how to handle meta-analyses with overlapping evidence—that is, the constituent meta-analyses were conducted on the same set or a substantial subset of the same primary studies. The approaches that have been taken to this nonindependence of evidence include simply ignoring the lack of independence, removing meta-analyses that are highly redundant with others, and conducting sensitivity analyses—that is, doing the second-order meta-analysis with different sets of constituent meta-analyses. Also, the ability of second-order meta-analyses to look at influences on the average effect sizes can be limited because the moderating and mediating variables examined must exist at the level of the meta-analyses that go into the second-order meta-analysis, not the individual studies. Still, second-order meta-analyses can be done (e.g., Tamim, Bernard, Borokhovski,

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Abrami, & Schmid, 2011). When the more desirable alternatives are not feasible, you should give consideration to doing a second-order meta-analysis.

EXERCISES

1. For the findings in the table below, what is the average weighted *d*-index?
2. Are the effect sizes of the seven studies homogeneous? Calculate your answer both by hand and by using a computer statistical package.

Finding	n_{i1}	n_{i2}	d_i
1	193	173	–.08
2	54	42	.35
3	120	160	.47
4	62	60	.00
5	70	84	.33
6	60	60	.41
7	72	72	–.28

NOTES

1. And they permit you to choose whether you want the test to estimate sampling error based on participant variation alone or both participant and study variation. I will return to this choice later, when I discuss fixed-effect and random-effects models.
2. Throughout this chapter and forward, I will use the terms *findings, studies,* and *comparisons* interchangeably to refer to the discrete, independent hypothesis tests or estimates of relationships that compose the input for a meta-analysis. I do this for exposition purposes, though sometimes these terms can have different meanings; for example, a study could contain more than one comparison between the same conditions.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

3. Borenstein et al. (2009) present formulas for how to combine two nonindependent effect sizes. These authors also provide formulas for how to combine effect sizes for different outcome measures taken on the same sample and for the same outcome measure taken on the same sample but at different times. The *d*-index for any two-group comparison can also be calculated if you have the means, samples sizes, and overall multi-degree-of-freedom *F*-test using the *Practical Meta-Analysis Effect Size Calculator* (Wilson, 2015).

4. Remember that measurement reliability can also be used as a moderator of effects, so without adjusting measures you could group them by reliability and ask, "Is the size of the impact of homework related to the reliability of the achievement measure?"

5. Half-standardizing is an alternative way to create similar slopes when only outcomes are dissimilar (see Greenwald, Hedges, & Laine, 1996).

6. The use of structural equation modeling in meta-analysis is an emerging area that incorporates many of the approaches I have described, not only to exploring multiple relationships in the same analysis, but also different model assumptions and even missing data techniques (Cheung, 2015). Synthesists will need a comfortable knowledge of these methods of structural equation modeling before they can use them successfully, though they can use the available software packages to carry them out.

Do not copy, post, or distribute

Copyright ©2017 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.