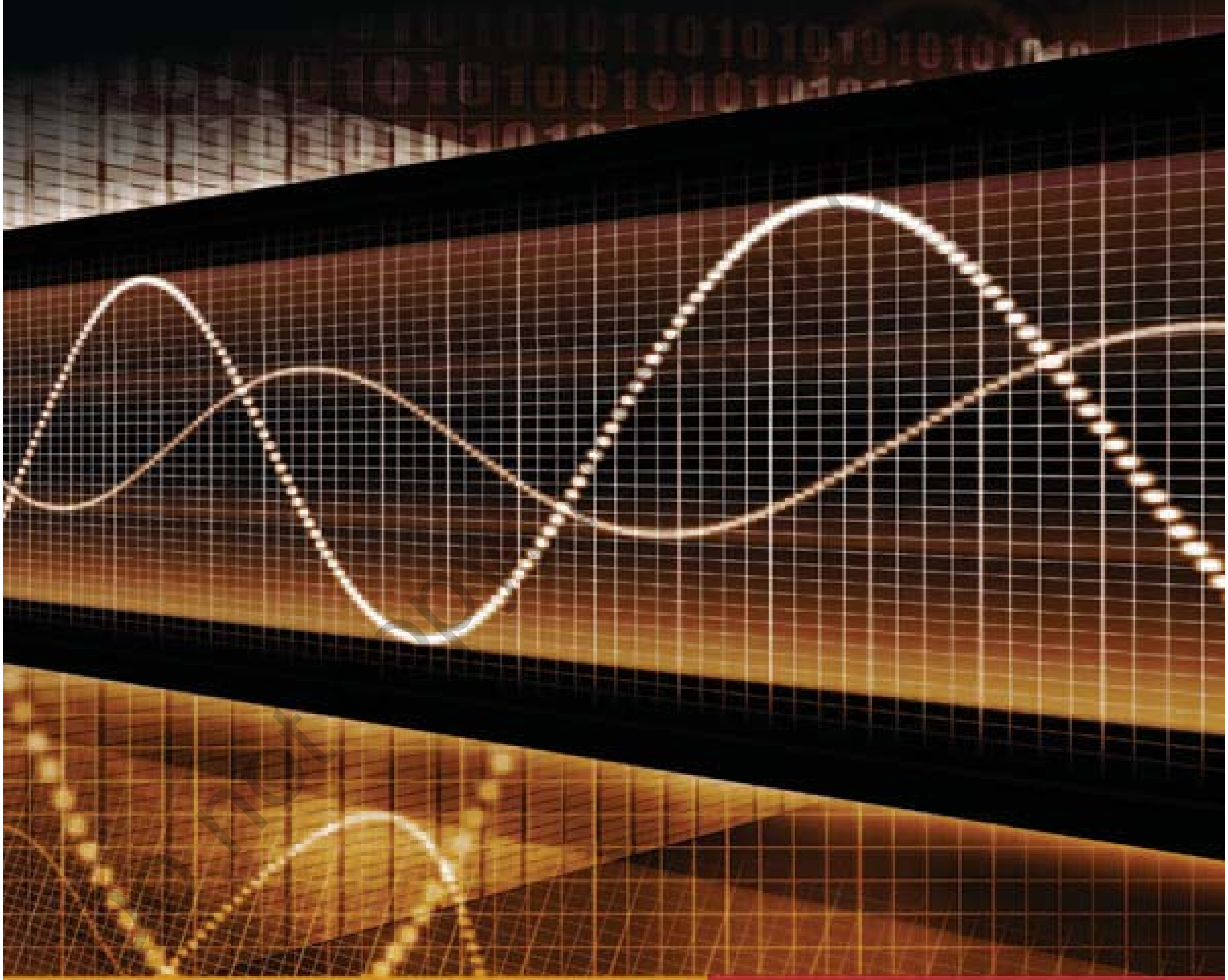
RESEARCH SYNTHESIS AND META-ANALYSIS

A Step-by-Step Approach



HARRIS COOPER

FIFTH EDITION

APPLIED SOCIAL RESEARCH METHODS SERIES

Volume 2

4

Step 3

Gathering Information From Studies

What procedures should be used to extract information from each study report?

Primary Functions Served in the Synthesis

1. To create a coding frame for obtaining information from studies
2. To train coders
3. To assess the accuracy of extracted information

Procedural Variation That Might Produce Differences in Conclusions

1. Variations in the information gathered from each study might lead to differences in what is tested as an influence on cumulative results.
2. Variations in coder training might lead to differences in entries on coding sheets.
3. Variation in rules for deciding what study results are independent tests of hypotheses might lead to differences in the amount and specificity of data used to draw cumulative conclusions.

Question to Ask When Evaluating the Information Gathered From Each Study to Be Included in a Research Synthesis

Were procedures used to ensure the unbiased retrieval of information from study reports?

This chapter describes

- How to construct a coding guide that will gather the important information about studies to be included in a research synthesis
- How to train coders so the information about studies will be gathered reliably
- Issues in judging whether separate outcomes from the same study should be considered independent outcomes
- What to do when information about a study is missing

So far, you have formulated the problem you want to explore in your research synthesis. You know the crucial issues that have come to the attention of theorists, researchers, and previous synthesists. And your literature search is underway. The next step in your synthesis is to begin the construction of your coding guide. The coding guide is the device you (and those who are assisting you) will use to gather information about each study. Most of this information will come from the study report itself, but some information may come from other sources as well.

INCLUSION AND EXCLUSION CRITERIA

I touched on how you make judgments about the relevance of studies when I discussed how a problem gets defined: you tie conceptual variables to observable research operations and measurements. Broadly defined concepts in a research synthesis will encompass more operational definitions than narrowly defined concepts. After the initial screening of studies, the coding guide you devise will direct the retrieval of information from studies. The guide needs to tell coders what characteristics of studies need to be present for a study to be included in the synthesis. It is where the conceptual rubber hits the operational road.

But conceptual relevance might not be the only criterion you wish to use for inclusion of studies. You might decide that a study that examines the hypothesis or intervention of interest to you conceptually does not match up with other criteria you want studies to meet. For example, you might want to limit studies based on when they were conducted. Our homework research synthesis excluded studies conducted before 1987. We used this criterion because an early synthesis ended with that year and we did not want our synthesis to cover overlapping research. The synthesis on aerobic exercise was limited to studies that used random assignment of participants to treatments. These were plentiful enough that limiting the synthesis to this type of research design, the one that allows the strongest causal inferences, was feasible (more on this in Chapter 5). In addition to timeframe and study design, other possible inclusion or exclusion criteria include characteristics of the study's context (e.g., its authors, dissemination outlet, funding source), participant sample (e.g., age, sex, economic status, geographic location), and outcomes (types of measures and their psychometric characteristics).

Sometimes inclusion and exclusion criteria other than conceptual relevance can be applied before the coding of studies even begins; it is easy to identify and exclude studies that are older than you wish. It is also possible, however, that you will begin coding studies and decide after the fact that additional screens need to be added. The coding sheet should allow you to do this. Also, you may decide that rather than exclude studies based on, say, the country in which they were conducted, you will use this variation in study context as a possible moderator of study outcomes.

DEVELOPING A CODING GUIDE

If the number of studies involved in your synthesis is small, it may not be necessary before you begin to examine the literature to have a precise and complete idea about what information to collect about the studies. The relevant reports, if only a dozen or so exist, can be retrieved, read, and reread until you have a good notion of what aspects of the studies would be interesting to code or how often the important characteristics suggested by others actually appear in the studies. For example, you might be interested in whether the effect of homework is moderated by the SES of students but you find that very few studies report the SES of the students taking part.

Of course, *if you read the entire literature first and then decide what information to code about each study, your choices of codes are post hoc and should not be solely dictated by what your reading suggested will be significant predictors of results*. If you do this, the proportion of significant results you get might be greater than if you chose predictors based solely on their theoretical or practical importance. Still, small sets of studies allow you to follow up on ideas that emerge only after the studies have been read. Then, you can return to previously read studies to code the new information you did not realize was important during the first reading.

If you expect to uncover a large number of studies, reading then rereading reports may be prohibitively time-consuming. In this case, it is necessary to consider carefully what data will be retrieved from each research report before the formal coding begins. Of course, reading a few randomly chosen studies can help you think about information to code and is something you should do. In fact, if you are interested in conducting a research synthesis, you are probably already familiar with many studies in the area.

When an area of research is large and complex, the construction of a coding guide can be no small task. The first draft of a coding guide should never be the last. First, you need to list all the characteristics of studies you want to gather. Then, you need to consider what possible values studies might take on each variable. For example, in a research synthesis of interventions to increase aerobic exercise among adults, you would certainly want to gather information on the age of participants and characteristics of the interventions, such as their length and intensity. You might decide that your definition of the term *adults* includes people over the age of 18, but participants still might be much older than this, which might influence the effects of the intervention. So you might want your coding guide to help you gather information on the range in age among participants. You might exclude studies involving adolescents, but the coding guide would still contain a question about the age of the youngest participant in the study, one about the oldest participant, and the mean and/or median age of participants.

After you have this preliminary set of coding questions and response categories, you need to show this first draft to knowledgeable colleagues for their input. They are certain to suggest additional codes and response categories. They will also point out instances in which your questions and responses are ambiguous and thus difficult to understand. After taking their advice, you should code a few randomly selected studies using the coding guide. This will add further precision to questions and response categories.

An important rule in constructing a coding guide for research synthesis is that *when many studies are involved, any information that might possibly be considered relevant should be retrieved from the studies.* Once data coding has begun, it is exceedingly difficult to retrieve new information from studies that have already been coded. Some of the information you gather on the coding sheets may never be examined in your completed synthesis. Sometimes, too few studies will report information about the variable of interest. In other cases, studies will not vary enough across values of a characteristic to allow valid inferences. For example, you might include a question about the health status of participants and discover that most if not all exercise interventions have been conducted with participants who have experienced a health problem. Still, it is much less of a problem to gather more information with your coding guide (by including a question about health status) that you may eventually find useful than it is to have to return to reports to get information that was neglected the first time through.

INFORMATION TO INCLUDE ON A CODING GUIDE

While the content of every research synthesis coding guide will be unique to the question asked, there are certain broad types of information that every synthesist will want to gather from primary research reports. Here, I will classify these types of information into eight categories:

1. The report
2. The predictor or independent variable
   a. If the report describes an experimental manipulation, information about the manipulated conditions—that is, the intervention (such as homework or exercise programs) or the independent variable (if the study is testing basic theoretical predictions, such as the effects of task choice)
   b. If the report describes nonmanipulated predictor variables, information about how these were collected and their psychometric characteristics (e.g., the scales used to measure participants' individual differences and attitudes toward rape)
3. The setting in which the study took place
4. Participant and sample characteristics
5. The dependent or outcome variables and how they were measured (such as level of achievement, amount of physical activity, motivation, or rape myth acceptance)
6. The type of research design
7. Statistical outcomes and effect sizes
8. Coder and coding process characteristics

In this chapter I will focus on six of the eight types of information about a study. I will return to discuss how to code research designs in Chapter 5 and statistical outcomes in Chapter 6, when each of these topics is covered in more detail.

A general coding guide will never capture all the important aspects of all studies. The questions that should guide your construction of the material to be retrieved from studies should include the following:

- Are there any theoretical and applied issues that need to be captured in the coding?
- Do theories suggest what study characteristics might be important and how the studies might differ on these characteristics?

- Are there issues in practical application that suggest that the way studies are conducted could relate to the impact of the intervention or policy?
- Are there any methodological issues that have arisen in the interpretation of past research?
- How might methods vary in ways that could relate to study outcomes?
- Are there disputes in the literature that relate to how studies are conducted?

Finally, completed coding sheets are often characterized by numerous entries left unfilled (I will return to this later) and notes in margins. Coders sometimes will feel as though they are slamming round pegs into square holes. Perfection is never achieved. Therefore, *it is good practice to leave coders space to make notes about on-the-spot decisions they made.* In general, the rules for constructing a coding guide are similar to rules used in creating a coding frame for a primary research effort (Bourque & Clark, 1992); a more-detailed description of the process in research synthesis can be found in Wilson (2009) and Orwin and Vevea (2009).

Characteristics of the report. Table 4.1 provides an example of a coding guide for report characteristics. My example is set up for coding on printed pages. If coders are coding directly into a spreadsheet, the response column is not needed. It is extremely important, however, that the spreadsheet clearly identify which column is devoted to which code. If you are using a program, such as Access, each coded variable can have its own page and coders can then click the appropriate response box. Also, when coding directly into spreadsheets you may forgo numbering many responses and simply type the coded response into the spreadsheet cell, for example by typing in "journal" rather than "1" for question R4 below.

Of course, if you type responses directly into the spreadsheet, spelling mistakes will appear as separate categories of response. On the positive side, typing in responses can be good for spotting errors. Typing "journal" into a spreadsheet column will make it obvious if the entry is in the wrong column, more so than if you are typing in numbers, which might be used repeatedly in adjacent columns.

Table 4.1 Example Coding Sheet for the Report Identification Section of a Coding Guide

Report Characteristics	
R1. What is the report ID number?	____ ____ ____
R2. What was the first author's last name? (Enter ? if you can't tell.)	____________
R3. What was the year of appearance of the report or publication? (Enter ? if you can't tell.)	__ __ __ __
R4. What type of report was this? 1 = Journal article 2 = Book or book chapter 3 = Dissertation 4 = MA thesis 5 = Private report 6 = Government report (federal, state, country, city) 7 = Conference paper 8 = Other (specify) ____________________ ? = Can't tell	____________
R5. Was this a peer-reviewed document? 0 = Not peer reviewed 1 = Peer reviewed ? = Can't tell	____________
R6. What type of organization produced this report? 1 = University (specify) __________ 2 = Government entity (specify) __________ 3 = Contract research firm (specify) ________ 4 = Other (specify) ____________________ ? = Can't tell	____________

(Continued)

Table 4.1 (Continued)

Report Characteristics	
R7. Was this research conducted using funds from a grant or other sponsor? 0 = No 1 = Yes ? = Can't tell	________
R7a. If yes, who was the funder? 1. Federal government (specify) ________ 2. Private foundation (specify) ________ 3. Other (specify) ________	________

Note that an "R" is placed before each question number in the first column. This was done to distinguish questions about the report from questions about other features of the study, which will be given other letters, such as "I" for intervention characteristics and "O" for outcome characteristics. Doing this is really a matter of personal taste: you could also just number the questions successively. Also, note that all possible responses to each question are listed below the question and each response is given a number that will be entered by the coder into the spaces provided in the second column. Some responses are simply "other." This code will be used if the coder finds a report characteristic that does not correspond to any response listed above it. When the "other" response is used, the coder is asked to provide a brief written description of the characteristic. Some of the questions also provide a "can't tell" response. Coders will use a question mark in the response space for "can't tell." This makes it easy to distinguish missing information from other coded values. I have repeated the "can't tell" response in most of the questions, but to save space coders could be instructed simply to use this convention throughout the coding sheet.

You will want to start by giving each report a unique identification number (Question R1). Later, you will also give unique numbers to each study in a report (if there is more than one study in it), to each unique sample within a study for which separate data are reported, and to each outcome reported within each sample.

Next, you will want to include on your coding sheet enough information about the first author of the study so that if you later want to group studies by their author (perhaps to test whether different authors get different results), you will be able to do so. In Table 4.1, the first author's name is used for this purpose (Question R2). Note that this is one of only two responses in the coding guide example that do not use numbers; the other one asks for the postal code of the state in which the study was conducted.

Third, you will want to know the year in which the report appeared. This might be used later to examine temporal trends in findings, or simply to help uniquely identify the study (along with the first author's name) in summary tables.

Fourth, you will want to describe the type of report and whether the report had undergone some form of peer review before it appeared. This information will later be used to test for the possibility of publication bias. Note here that the response categories are "mutually exclusive" (every report should fall into only one category) and "exhaustive" (every report will have a category).

Finally, you might be interested in what type of organization produced the report and whether the report was done with some type of funding support. This information can be critical if you discover that the funders of some studies might have had a monetary or other interest in whether studies had a particular outcome; an example would be a chain of gymnasiums that supports a study on the value of exercise for older adults. If so, you might want to see if such studies produce different results from unfunded studies. The importance of gathering this information will depend on your research problem.

Experimental conditions, if any. You will need to describe carefully the details of any experimental conditions—that is, the intervention or independent variable—if these were part of the study. This portion of the coding guide describes the relevant operations that define the experimental conditions and the categories that capture the variations in how the conditions might have been operationalized. What was experienced by people in the experimental condition? What was the intensity and duration of the intervention? As important as it is to describe what happened in the experimental condition, it is equally important to describe how the control or comparison group was treated. Was there an alternate intervention? If so, what was it? If not, what did participants in the comparison conditions do, or how were controls obtained? Differences among studies on any of these variables would be prime candidates for causes of differences in study outcomes.

Table 4.2 provides some examples of the types of information that might be gathered on a coding sheet for studies comparing students

Table 4.2 Example Coding Sheet for Homework Interventions (Selected Questions)

Information About the Homework Intervention (Complete these questions separately for each homework intervention described in the study report.)	
I1. What is this study's ID number?	____________
I2. Which of the following characteristics were part of the homework intervention? (Place a 1 in each column that applies, 0 if not, ? if not reported.)	(Page found__)
1 = Focuses on academic work	____________
2 = Assigned by classroom teachers (or researcher via teacher)	____________
3 = Meant to be done during nonschool hours or during study time at school	____________
I3. What was the subject matter of assignments? (Place a 1 in each column that applies, 0 otherwise)	
1 = Reading	____________
2 = Other language arts	____________
3 = Math	____________
4 = Science	____________
5 = Social studies	____________
6 = Foreign language	____________
7 = Other (specify) ________________________	____________

I4. How many homework assignments were assigned per week? (Enter ? if not reported.)	____________
I5. What was the expected amount of time needed to complete each assignment, in minutes? (Enter ? if not reported.)	____________
I6. Were assignments graded? 0 = No 1 = Yes ? = Can't tell	____________
I7. Was the homework used in determining class grades? 0 = No 1 = Yes ? = Can't tell	____________
I8. Was evidence reported that the homework intervention was or was not implemented in a manner similar to the way it was defined? (An example of when you would answer "not implemented as specified" to this question would be if the report says the homework was meant to be assigned three times a week but was assigned only once a week.) 0 = Not implemented as specified: What information was used to make this determination? ________________________________ 1 = Implemented as specified: What information was used to make this determination? ________________________________ ? = Nothing reported about the fidelity of implementation	____________

(Continued)

Table 4.2 (Continued)

I9. Was there evidence that the group receiving homework might also have experienced a changed expectancy, novelty, and/or disruption effect that the control group did not also experience? 0 = No change in expectancy, etc. 1 = Yes, change in expectancy, etc. ? = Nothing reported about change in expectancy, etc.	____________
I10. How was the comparison group treated? 0 = No homework and no other compensating activity 1 = Other compensating activity (specify) ________________________________ ? = Not reported	____________
I11. Were the homework and comparison group drawn from the same school building and were they in the same grade? 0 = No 1 = Yes ? = Not reported	____________
I11a. If yes to I11, did the students, parents, and/or teachers in either the homework or comparison group know who was in which condition? 0 = No 1 = Yes ? = Not reported	____________

who did homework with students who did not do homework. First, note that the homework intervention code sheet gives each intervention described within the report a unique study number. This allows for the possibility that there might be more than one study of homework described in a single report, or that there might be more than one homework intervention within the same study (e.g., some students did an hour of homework, other students did a half hour, and still others did no homework at all).

Note that the second cell of the second column of the coding sheet (Question I2) now also asks coders to give the page number on which the information was found in the report. This is a good procedure and speaks in favor of using print coding guides. This is an excellent procedure to follow when the placement of information in the report might not be clear. (I did not do this in Table 4.1 because all the information should be available on the report's title page or front matter.) Later, if coders have concerns about how they coded particular pieces of information or if two coders disagree about a code, having its location reported on the coding sheet will ease the process of finding the information for checking and can save much time. To save space, I have shown this only once in the tables, but it can appear for just about every question. If coders are working from their own copies of reports, you can also ask them to circle or highlight (in a pdf) the place in the report where information was found and put the question number from the coding guide on the report as well. It will then be easy to see where each code came from.

The next question asks whether this homework intervention meets each of three characteristics that define homework. If any of these three are answered "no," it might lead to the study being excluded from the analysis. The next five questions ask about other characteristics of assignments that the synthesists might want to test as moderators of the effects of the homework intervention. There might be more of these. Note that Question I3 (i.e., "What was the subject matter of assignments?") uses numbers to distinguish seven different coded responses and each is given a "applies," or "does not apply), or "not reported" answer. The reason for this is that a homework assignment might cover any one of the six subject matters or any combination of two or more. There are dozens of such combinations. It would be tedious for you and coders if you listed them all out, especially since you know that most of the combinations would never be coded. By coding the subject matters using just these seven codes (which still give you precise information about each study), you can then examine how frequently each combination occurs and have the computer create new variables based on the codes. For example, you might find that most studies cover only one subject matter but a few cover both reading and language arts. So, you could instruct the computer to create a new variable that has eight values, one for each instance in which only one subject matter gets a 1 and the others get a 0, a seventh in which both reading and language arts get a 1, and an eighth for all other combinations.

Question (I8) relates to the fidelity with which the homework intervention was carried out. If the way homework was actually carried out in studies was different from the intended treatment, this might raise questions about whether the study was a fair test of homework's effects. For this question, the coding sheet also provides a note that is meant to help the coder remember a coding convention that was established to clarify how to code a study. In this case, the meaning of "implemented in a similar manner" might be ambiguous, so the note clarifies its meaning. Using these notes will help ensure that different coders use the guide in the same manner, thus reducing differences between them (and also within a particular coder's responses from study to study). The next question (I9) asks whether there was evidence that the homework intervention was confounded with other differences in the way the experimental and control group were treated. If such confounds exist it would compromise the study's ability to draw causal inferences about the effects of homework. Answers to either or both of these questions might lead to the study being excluded from the synthesis, or the information might be used to group studies to see if these characteristics were associated with study outcomes.

Questions I10 and I11 relate to the control participants. Question I10 asks about how the control group was treated and Question I11 tries to get at whether the participants in each condition knew there were other participants in the study who were being treated differently. If so, this might have influenced how they behaved. Each of these questions (I10–I11) relates to the construct validity of the treatment manipulation.

Setting of the study. This information most often includes the geographic location of the study (e.g., country, state, or part of the country; urban, suburban, or rural community). If the studies have been conducted within an institutional setting—for example, schools, hospitals, or gymnasiums—this information could be gathered as well. Furthermore, some studies will always be conducted within institutional contexts (e.g., homework studies always occur within schools), so differences in institutions might be of interest (e.g., "Was it a public or private school?," "Did the school have a religious affiliation?," etc.). Table 4.3 presents some example questions related to setting that might appear on a homework code sheet.

Participants and samples. Another type of information typically collected from research reports concerns the characteristics of the participants included in the primary research. This can include the age, race and/or ethnic group, and social class of participants, as well as any restrictions placed by the primary researchers regarding who could participate in the study. Table 4.4 provides some examples of participant and sample characteristics that might be important in a study of the effects of homework. Also, note that yet another unique sample ID number must be provided here because some studies might present information on separate samples within the study. For example, a study might break out its samples and results based on whether students were high achieving or average. To capture this distinction, each sample would get a different sample number and Question P2 would be answered differently for each sample.

Predictor and outcome characteristics. For studies that do not involve experimental manipulations but rather associate measured variables, one as the predictor of another (e.g., individual differences predicting rape attitudes), or for the outcomes of studies with experimental manipulations (e.g., measures of cognitive functioning by adults after aerobic exercise, or motivation after a choice manipulation), you will want to retrieve information concerning the types of outcomes and whether they were standardized measures, and evidence about the outcomes' validity or reliability, if this information is available.

Table 4.3 Example Coding Sheet for Study Setting Characteristics in a Homework Synthesis

Setting Characteristics	
S1. Where were the participants? (Place a 1 in each column that applies, 0 if not, ? if not reported.) 1 = In the United States 2 = In a country other than the United States (specify country) ________________________________	 ________
S2. What state(s) was the study conducted in? (Use postal service code/s.)	________
S3. What type of community was the study conducted in? 1 = Urban 2 = Suburban 3 = Rural ? = Can't tell	________
S4. What type of school was the study conducted in? 1 = Public school 2 = Private school (secular) 3 = Private school with a religious affiliation (specify religious group) ____________________ ? = Can't tell	________
S5. What classroom types were represented among the settings? (Place a 1 in each column that applies, 0 if not, ? if not reported.) 1 = Regular education 2 = Special education 3 = Other (specify) ____________________ 4 = No classroom types given	 ________ ________ ________ ________

Table 4.5 provides some questions that might be asked about the outcomes of a homework study. Note that the first question requires, yet again, that a unique number be given to each outcome. So, we now have a four-tiered system that, when the ID numbers

Table 4.4 Example Coding Sheet for Participant and Sample Characteristics in a Homework Synthesis

Participant and Sample Characteristics (Complete these questions separately for each sample within a homework intervention comparison for which there is a separate outcome.)	
P1. What is this sample's ID number?	________
P2. Which of the following labels were applied to students in this sample? (Place a 1 in each column that applies, 0 if not, ? if not reported.) 1 = High achieving 2 = Average 3 = "At risk" 4 = Underachieving/below grade level 5 = Possessing a learning deficit 6 = Other (specify) ____________	 ________ ________ ________ ________ ________ ________
P3. What was the SES of students in the sample? (Place a 1 in each column that applies, 0 if not, ? if not reported.) 1 = Low SES 2 = Low-middle SES 3 = Middle SES 4 = Middle-upper SES 5 = Upper SES 6 = Only labeled as *mixed*	 ________ ________ ________ ________ ________ ________
P5. What were the grade levels of the students in the sample? (Place a 1 in each column that applies, 0 if not. Use options 13 through 16 only if no specific grade information was reported.) 0 = K 1 = 1 2 = 2 3 = 3	 ________ ________ ________ ________

(Continued)

Table 4.4 (Continued)

4 = 4	______________
5 = 5	______________
6 = 6	______________
7 = 7	______________
8 = 8	______________
9 = 9	______________
10 = 10	______________
11 = 11	______________
12 = 12	______________
13 = Labeled as *elementary school*	______________
14 = Labeled as *middle school*	______________
15 = Labeled as *junior high school*	______________
16 = Labeled as *high school*	______________
17 = No grade level information given	______________
P6. What sexes were represented in the sample? (Place a 1 in each column that applies, 0 if not.)	
1 = Males	______________
2 = Females	______________
3 = No sex information given	______________
P6a. If reported, what was the percentage of females in the sample? (Use ? if not reported.)	______________

are strung together, uniquely identifies each outcome within each sample, each sample within each study, and each study within each report. In some studies outcomes will be reported for, say, more than one grade level or more than one measure of achievement. When such a study is uncovered, the coder would fill out separate sheets for each two-group combination. For example, a study with both a standardized test and a class grade measure of achievement reported separately for students in fifth grade and sixth grade would have four outcome coding sheets associated with it, two each for fifth and sixth graders.

Table 4.5 Example Coding Sheet for Outcomes in a Homework Synthesis

Outcome Measure (Complete these questions separately for each relevant outcome within each sample.)	
O1. What is this outcome's ID number?	____________
O2. What subject matter did this outcome measure? (Place a 1 in each column that applies, 0 if not.) 1 = Reading 2 = Other language arts 3 = Math 4 = Science 5 = Social studies 6 = Foreign language 7 = Other (specify) ________________________ 8 = Not a subject matter test	 ____________ ____________ ____________ ____________ ____________ ____________ ____________ ____________
O3. What type of outcome measure is this? 1 = Standardized achievement test (specify) ____ 2 = Another test measuring achievement (e.g., teacher-developed, textbook chapter tests) 3 = Class grades after homework 4 = Multiple types of student achievement measures combined into one measure 5 = Student study habits and skills ? = Can't tell	____________
O4. Was evidence presented regarding whether the validity/reliability of this outcome measure reached an acceptable criterion? (Note: Place a 1 in each column if acceptable, 0 if not, ? if not reported. A statement indicating that internal consistency was	

(Continued)

Table 4.5 (Continued)

"acceptable" is sufficient, even if the specific value was not reported. A citation to an external source is sufficient.) 1 = Internal consistency 2 = Test-retest correlation 3 = Other (specify) ________________________	 __________ __________ __________
O5. How many days after the homework intervention was the outcome measure administered? (Enter 0 if outcome measure was given on the last day of the homework study. Enter ? if unable to determine.)	__________

Note as well that it is not just different measures of the same construct that can create multiple measures associated with the same sample (within the same study within the same report). It is also possible for researchers to collect the same measure two or more times. That is one reason why Question O5 is included on the outcome code sheet. Also, researchers might collect data on more than one construct. For example, the homework synthesis might not have focused exclusively on achievement but might also have collected outcomes related to study skills and/or attitudes toward school. If this were the case, the outcomes coding sheets would be expanded to include questions and responses related to measures of these constructs.

The fourth question (O4) on the outcome code sheet relates to the validity and reliability of the measure. These questions can be phrased in lots of different ways, depending on the level of detail you wish to gather. The example requests information that is not very specific, asking the coders only whether the measure reached an "acceptable" level of reliability.

Coder and coding characteristics. The coding guide should contain a section for the coders to enter their names or ID number and the

Table 4.6 Example Coding Sheet for Coder and Coding Information

Coder and Coding Characteristics	
C1. What is your coder ID number?	____________
C2. On what date did you complete coding this study?	___/___/___
C3. In minutes, how long did it take you to code this study?	___ ___ ___
Notes (provide below any notes about the study or concerns you had regarding your codes):	

date on which they coded the study (see Table 4.6). You might also ask coders to state the amount of time it took for them to code the study, for accounting purposes. In some instances, this information might be formally incorporated into your data files. This section can also provide coders with space to make any narrative comments about the coding process they want to share with you.

Low- and High-Inference Codes

Most of the information requested in the example coding guides might be thought of as low-inference codes. That is, they require the coder to locate the needed information in the research report and transfer it to the coding sheet. In some circumstances, coders might be asked to make some high-inference codes about the studies. It might have occurred to you that there were some inferences that coders were asked to make on the homework coding sheets. For example, I noted previously that coders using the example guide for outcomes (Table 4.5) would be asked to code whether the estimates attained for the internal consistency, test-retest reliability, and other validity/reliability estimates for measures were "adequate" (Question O4). If left to their own devices, the judgment of adequacy would indeed be a somewhat subjective judgment, one that might vary from coder to coder. However, if you gave coders a threshold that defined "adequate," the need for judgment would have been removed from these questions. So, the question might have been rephrased to ask, "Was an estimate of internal consistency present? If yes, was it above .8?" Or the coders might have been asked to gather the exact values of the internal consistency estimates. The exact values then could be used to test whether this measure of the validity/reliability of the measures was related to study outcomes.

Other high-inference codes involve attempting to infer how an intervention or experimental manipulation might have been experienced by the individuals presented with it. A synthesis by Carlson and Miller (1987) provides a good example. They summarized the literature on why negative mood states seem to enhance the likelihood that people will lend a helping hand. In order to test different interpretations of this research, they needed to estimate how sad, guilty, angry, or frustrated different experimental procedures might have made participants feel. To do this, a group of judges were asked to read excerpts from the methods sections of relevant articles. The judges then used a 1 to 9 scale to rate the "extent to which subjects feel specifically downcast, sad, or depressed as a result of the negative-mood induction" (p. 96). These judgments were then added to the coding sheets for each study.

These high-inference codes create a special set of problems for research synthesists. First, careful attention must be paid to the reliability of high-inference judgments. Also, judges are being asked to play the role of a research participant, and the validity of role-playing methodologies has been the source of much controversy (Greenberg & Folger, 1988). However, Miller, Lee, and Carlson (1991) empirically demonstrated that high-inference codes can lead to valid judgments and can add a new dimension to synthesists' ability to interpret literatures and resolve controversies. This technique deserves a try if you believe you can validly extract high-inference information from articles and persuasively explain your rationale for doing so (i.e., how it will increase the value of your synthesis).

SELECTING AND TRAINING CODERS

The coding of studies for a research synthesis is not a one-person job. Even if a single person eventually does gather information from all the studies, the research synthesists must demonstrate that this person did a good job of data extraction. There is simply too much room for bias (conscious or unconscious), for idiosyncratic interpretation of coding questions and responses, and for simple mechanical error for the unverified codes of a single person to be considered part of a scientific synthesis of research. For example, Rosenthal (1978) looked at 21 studies that examined the frequency and distribution of recording errors. These studies uncovered error rates ranging from 0% to 4.2% of all the data recorded; 64% of the errors in recording were in a direction that tended to confirm the study's initial hypothesis (see also Leong & Austin, 2006).

Recording errors are not the only source of unreliability in study coding. Sometimes, codes cannot be reliably applied because the reports of studies are not clear. Other times, ambiguous definitions provided by the research synthesists lead to disagreement about the proper code for a study characteristic. Finally, as I noted earlier, the predispositions of coders can lead them to favor one interpretation of an ambiguous code over another.

Stock and colleagues (Stock, Okun, Haring, Miller, & Kinney, 1982) empirically examined the number of unreliable codings made in a research synthesis. They had three coders (one statistician, and two post-Ph.D. education researchers) record data from 30 documents into 27 different coding categories. Stock and colleagues found that some variables, such as the means and standard deviations of the ages of participants (a low-inference code), were coded with perfect or near-perfect agreement. Only one judgment, concerning the type of sampling procedure used by the researchers, did not reach an average coder agreement of 80%.

Demonstrating that the coding definitions are clear enough to generate consistent data across coders and that the coders have extracted information from the reports accurately—that is, gave responses to the coding questions that were little different from those that would have been given by any other coder—will involve training at least two coders. Doing so is especially important if the number of studies to be coded is large or if persons with limited research training are called on to do the coding. It is rare today to find a research synthesis in which a single coder gathered information from all studies—and any such syntheses are looked on skeptically. Most syntheses involve at least two coders gathering information from at least a portion of the studies. Some syntheses involve teams of three or more coders. In any case, *it is good practice to treat the coding of studies as if it were a standard exercise in data gathering.*

Some synthesists will have every study coded independently by more than one coder, called *double coding*. The codes for every study are then compared, and discrepancies are resolved in a meeting of the coders or by a third party. This procedure can greatly reduce potential bias, make evident different interpretations of questions and responses, and catch mechanical errors.

While all synthesists must demonstrate the reliability of their codes, how far they can go to ensure reliability will be a function of the number of studies to be coded, the length and complexity of the coding guide, and the resources available to accomplish the task. Clearly, syntheses involving larger numbers of studies with complex coding sheets will require more coding time. Unless lots of time is available, more studies to code will make it more difficult to have every study coded twice. In some cases, if there is complex information to be coded, synthesists can decide to double code some of the information on the coding sheet but not other information. The synthesists must determine how to get the most trustworthy codes possible given their limited resources.

Double coding is not the only way you can enhance the reliability of codes. First, you can pick coders who have the background and interest needed to do a good job. People with lots of experience reading and conducting research make better coders than novices. Training can overcome some limitations of inexperience, but not all.

Second, coding sheets can be accompanied by coding guides that define and explain distinctions in each study characteristic. In the examples given in Tables 4.1 through 4.6, some of these definitions appear directly on the coding sheet. A coding guide with other definitions and conventions for coding particular questions could accompany the coding sheets. The more, the better.

Third, prior to actual coding, discussions and practice examples should be worked through with coders. *It is important to pilot test your coding guide using the individuals who will actually do the coding*. Use a few research reports, preferably chosen to represent what you know are diverse types of research contained in the literature, and talk through how the coding would proceed. The coders will raise concerns you had not thought of, which will lead to greater clarity in questions, responses, and conventions to use when reports are unclear.

Fourth, the coders should gather information for the same few studies independently and share their responses in a group. You should discuss mistakes with them. Even-greater clarity in the coding guide will result. At this stage and during subsequent coding, some synthesists will attempt to keep the coders unaware of certain aspects of the studies. Some will remove information about the study's authors and affiliation from the report so that coders will not be influenced by any knowledge they may have about the researchers. Some synthesists have the different sections of the report coded by different coders so that, for example, the results of a study do not influence the ratings it might get on the quality of the study design. These procedures are more important to follow when (a) coding decisions might involve high-inference judgments, (b) the research area is distinguished by polarized opinions and findings, and/or (c) the coders are themselves very knowledgeable about the area and might have their own opinions about what the results of studies "should" be.

Estimating reliability. Once these steps have been completed, you are ready to assess reliability. This should happen before coders are given lots of studies to code and again periodically during coding. It is usually important to obtain numerical estimates of coder reliability. There are many ways to quantify coder reliability and it appears that none is without problems (see Orwin & Vevea, 2009, for a general review of evaluating coding decisions). Two methods appear most often in research syntheses. *Most simply, research synthesists will report the agreement rate between pairs of coders.* The agreement rate is the number of agreed-on codes divided by the total number of coding opportunities. Typically, the percentage of agreement will be broken out by each coding question. If the number of codes is large, the synthesists may provide only the range of agreement percentages and then discuss any that might seem problematically low. For example, in the synthesis of studies relating choice to intrinsic motivation, we found that out of a total of 8,895 codes, there were 256 disagreements; that is, coders disagreed 2.88% of the time. The question that gave coders the most trouble involved the description of the control group, with disagreements occurring 9.4% of the time for this variable.

Also useful is Cohen's kappa, a measure of reliability that adjusts for the chance rate of agreement. The value of *kappa* is defined as the improvement over chance reached by the coders. Often kappa is presented along with the percent agreement.

As mentioned previously, some synthesists will have each study examined by two coders, will compare codes, and then will have discrepancies resolved through discussion or by consulting a third coder. This procedure leads to very high reliability; if it is used, it often is not accompanied by a quantitative estimate of reliability. In order to get an effective reliability for double coding, you would have to form two teams of two coders and an arbiter and compare the results of the two teams' deliberations. You can see that this process is unlikely to result in many differences between the teams, as long as the coding definitions are clear.

Other synthesists have individual coders mark the codes they are least confident about and discuss these codes in group meetings. This procedure also leads to highly trustworthy codes. Regardless of what techniques are used, the question to ask when evaluating the methods of data collection used to carry out research syntheses is,

> Were procedures used to ensure the unbiased and reliable (a) application of criteria to determine the substantive relevance of studies, and (b) retrieval of information from study reports?

Transferring Information to the Data File

In the foregoing paragraphs, I describe techniques for ensuring that information about each study was correctly recorded into coding sheets. I suggest that the best way to do this is to have each study coded by more than one researcher and then compare their codes to one another. Even if the coders agree on the coding sheets, *it is good practice to have two people transfer the results from the coding sheets into separate data files—the files that will be used by the computer when the data are analyzed.* Then, these files can be compared to one another to determine if any errors have been made when data were transferred from the coding sheets or placed directly into the computer. If only one coder is used, this person can be asked to do the data entry twice. Although this task may seem simple, errors in data transcription are to be expected, especially when the task data are complex. Of course, if a computer program such as Access is used that transfers codes directly into a data set ready for the computer, this type of check is unnecessary. However, the entries into Access still need to be checked.

PROBLEMS IN GATHERING DATA FROM STUDY REPORTS

In Chapter 3 I discussed some deficiencies in study retrieval that will frustrate synthesists regardless of how thorough and careful they try to be. Among these, some potentially relevant studies do not become public and defy the grasp of even the most conscientious search procedures. Other studies you will learn about but will not be able to obtain.

Perhaps the most frustrating occurrence in collecting the evidence is when synthesists obtain primary research reports but the reports do not contain the needed information. Reports could be missing information on study characteristics, preventing the determination of whether study outcomes were related to how the study was conducted, or even whether the study was relevant at all. Or information could be missing on statistical outcomes, preventing synthesists from estimating the magnitude of the difference between two groups or the relationship between two variables.

Imprecise Research Reports

Incomplete reporting will be of most concern to research synthesists who intend to perform meta-analyses. What should the meta-analyst do about missing data? Several conventions can be suggested to handle the most common problems.

Incomplete reporting of statistical outcomes. Research reports sometimes lack important information about the results of statistical procedures carried out by the primary researchers. Statistical data are often omitted when the researcher was testing to reject the null hypothesis and it is not rejected. Instead of giving the exact results of the statistical test, the researchers simply say it did not reach statistical significance. In these cases, the researchers are also less likely to provide the correlation or means and standard deviation associated with the finding. Sometimes they do not even tell which direction the correlation or comparison of group means was in.

You have limited options when you know a relationship or comparison has been tested but the primary researchers do not provide the associated means and standard deviations, sample size, inference test value, *p*-level, or effect size. One option is to contact the researchers and request the information. As I noted in Chapter 3, the success of this tactic will depend partly on whether the researchers can be located as well as on the status of the requester. The likelihood of compliance with the request will also depend on how easy it is for the researchers to retrieve the information. There is less chance a request will be fulfilled if the study is old, if the desired analyses are different from those originally conducted, or if the requester asks for a lot of data.

The chance of getting a response from researchers will increase if you can make the request as easy to fulfill as possible. This might include providing the researchers with a table in which they simply need to plug in the values you need. Never ask for more information or more detailed information than you need. The more information you ask for, the more authors may worry that you think they did something wrong and suspect that you are interested in more than just including their study in a meta-analysis. (Of course, it is also important to follow up with authors if you think you have uncovered an erroneous result.)

Another approach to finding missing data is to examine other documents that describe the study being reported. For example, if you have found a journal article that reports some but not all the results you need, but the accompanying Author Note says the study was conducted as a doctoral dissertation, it might be that the dissertation itself contains the information. Often, dissertations have appendixes that include thorough descriptions of results. Or some research reports prepared by government agencies and contract research firms might be written with audiences in mind who will not be interested in the details. These organizations also might have available more technical reports with lots more information in them.

If you cannot retrieve the needed data, another option is to treat the outcome as having uncovered an exact null result. That is, for any statistical analysis involving the missing data, a correlation of 0 is assumed, or the means being compared are assumed to be exactly equal. It is reasonable to expect that this convention has a conservative impact on the results of the meta-analysis. In general, when this convention is used, the cumulative average relationship strength will be closer to zero than if the exact results of nonsignificant relationships were known. However, adding zeros to your data set for missing values will change the characteristics of your distribution of findings. For these reasons, it is rare for meta-analysts to use this procedure anymore.

A fourth option is simply to leave the comparison out of your meta-analysis. This strategy will likely lead to a higher average cumulative relationship than if the missing value was known. All else being equal, nonsignificant findings will be associated with the smaller relationship estimates in a distribution of sampled estimates. However, most meta-analysts choose this fourth option, especially if the number of missing values is small relative to the number of known values. Also, if meta-analysts can classify missing value outcomes according to the direction of their findings—that is, if they know which group had the higher mean or whether the correlation was positive or negative—these outcomes can be included in vote count procedures (discussed in Chapter 6). It is possible to estimate the strength of a relationship using vote counts (see Bushman & Wang, 2009). Also, in Chapter 7 I will discuss ways to test meta-analytic results to see whether the conclusions would be different using different methods to handle missing data. When statisticians analyze the same data using different statistical assumptions, it is called *sensitivity analysis* (see Chapter 7).

Incomplete reporting of other study characteristics. Research reports also can be missing information concerning the details of study characteristics other than their outcomes. For example, reports might be missing information on the composition of samples (e.g., in a homework study the students' economic background), the setting (e.g., whether the school was in an urban, suburban, or rural community), or treatment characteristics (e.g., the number of homework assignments each week and their length). Meta-analysts want this information so they can examine whether treatment effects or relationship magnitudes are associated with the conditions under which the study was conducted.

You have several options when study information of this sort is missing. First, you can ask yourself whether the information might be available in sources other than the research report. For example, the homework coding guide contains a question about whether the school was in an urban, suburban, or rural community, and a question on the students' economic status. If you know the school district in which this study took place, this information might be available on the district or state website. If information on the psychometric characteristics of measures is not reported, these might be found in reports on the instruments themselves.

Most simply, you can leave the study with missing information out of the analysis, although it may be included in other analyses for which the needed information is available. For example, homework studies missing information on the students' economic background (a frequent occurrence) simply cannot be used in the analyses testing whether this characteristic influences the effect of homework, but they can be used in analyses looking at grade level, a characteristic rarely missing from reports.

Alternatively, it is sometimes appropriate to assume that a missing value suggests what the value is. This will happen because the researchers have assumed readers will take the information for granted. For example, homework researchers are likely in nearly all instances to mention if a study was conducted in an all-boys or all-girls school. So, when the sex composition of classes is not mentioned, it is probably safe to assume that both boys and girls were present, and perhaps in roughly equal numbers. You might have coders use "?" for this code but then have the computer consider this code to mean "both boys and girls." If you do this, you should mention the convention in your methods section when you write up your synthesis. Also, if possible, you might run this analysis twice, once with the studies coded "?" included and once without.

The amount of concern a meta-analyst should have over missing study characteristics will depend partly on why the data are missing. Some data will be completely missing at random. That is, there will be no systematic reason why some reports include information on the characteristic while others do not. If this is the case, then the outcome of an analysis examining the relationship between study outcomes and study characteristics will be unaffected by the missing data except, of course, for a loss of statistical power.

If the reason data are missing relates systematically to study outcomes, or to the values of the missing data themselves, then the problem is more serious. In this case, the missing data might be affecting the results of the analysis. For example, suppose health researchers are more likely to report that the participants in their study were all females or all males if the result indicates a significant effect of an activity intervention. Nonsignificant effects are more often associated with mixed-sex samples, but this is unknown to the meta-analyst because researchers who find nonsignificant results are less inclined to report the sample's composition. In such a case, the meta-analyst would have a hard time discovering the relationship between the sex composition of the intervention study and the magnitude of the intervention's effect (e.g., exercise is more or less effective when groups are composed of the same sex).

Pigott (2009) suggests several other strategies for dealing with missing study characteristics. First, missing values can be filled in with the mean of all known values on the characteristic of interest. This strategy does not affect the mean outcome of the cumulative analysis, except to raise its power. It is most appropriate when the meta-analyst is examining several study characteristics together in one analysis. In such a case, a single missing value may delete the entire study, which might not be desirable. Second, the missing value can be predicted using regression analysis. In essence, this strategy uses known values of the missing variable found in other studies to predict the most likely value for the missing data point. Pigott (2009) describes several more-complicated ways to estimate missing data.

In most instances, I would advise meta-analysts to stick with the simpler techniques for handling missing data. As techniques become more complex, more assumptions are needed to justify them. Also, when more-complicated techniques are used, it becomes more important to conduct sensitivity analysis. It is always good to compare results using filled-in missing values with results obtained when missing values are simply omitted from the analysis.

IDENTIFYING INDEPENDENT COMPARISONS

Another important decision that must be made when data are being gathered involves how to identify independent estimates of relationship strength or group differences. Sometimes a single study may contain multiple tests of the same comparison or relation. This can happen for several reasons. First, more than one measure of the same construct might be used by the researchers with measures analyzed separately. For example, a researcher of choice effects might measure intrinsic motivation using both participants' self-reports and observations of their activities during a free-play period. Second, measures of different constructs might be taken, such as several different personality variables all related to attitudes toward rape. Third, the same measure might be taken at two or more different times. And finally, people in the same study might be broken out into different samples and their data analyzed separately. This would occur, for instance, if a rape-attitude researcher gave the same measures to all participants but then separately examined results for males and females. In all these cases, the separate estimates in the same study are not completely independent—they share methodological and situational influences. In the case of the same measure taken at different times, the study results even share influences contributed by having been collected on the same people with the same measures.

The problem of nonindependence of study results can be taken even farther. Sometimes a single research report can describe more than one study conducted sequentially by the same research team in the same location. So, the two studies likely were conducted in the same context (e.g., the same laboratory), perhaps with the same research assistants, and with participants drawn from the same participant pool. Also, multiple research reports in the same synthesis often describe studies conducted by the same principal investigators. The synthesists might conclude that studies conducted by the same researchers at the same site, even if they appear in separate reports over a number of years, nevertheless contain certain constancies that imply the results are not completely independent. The same primary researcher with the same predispositions may have used the same laboratory rooms while drawing participants from the same population.

Synthesists must decide when statistical results will be considered independent tests of the problem under investigation. Several alternatives can be suggested regarding the proper unit of analysis in research syntheses.

Research Teams as Units

The most conservative way to identify independent results uses the laboratory or researcher as the smallest unit of analysis. Advocates of this approach would argue that the information value of repeated studies by the same research team is not as great as an equal number of studies reported from separate teams. This approach requires the synthesists to gather all studies done by the same research team and to come to some overall conclusion concerning the results for that particular group of researchers. Therefore, one drawback is that this approach requires the synthesists to conduct syntheses within syntheses, since decisions about how to cumulate results first must be made within research teams and then again between teams.

The research-team-as-unit approach is rarely used in practice. It is generally considered too conservative and too wasteful of information that can be obtained by examining the variations in results from study to study, even within the same laboratory. Also, it is possible to ascertain whether research teams are associated with systematic differences in study outcomes by using the researchers as a study characteristic in the search for outcome moderators.

Studies as Units

Using the study as the unit of analysis requires the synthesists to make an overall decision about the results reported in an individual study. If a single study contains information on more than one test of the same group comparison or relation, the synthesists can calculate the average of these results and have that represent the study. Alternatively, the median result can be used. Or if there is a preferred type of measurement—for example, a particular rape-attitude scale with good measurement characteristics—this result can represent the study.

Using the study as the unit of analysis ensures that each study contributes equally to the overall synthesis result. For example, a study estimating the relationship between rape attitudes and need for power using two different attitude scales and reporting separately for men and women would report four nonindependent correlations. Cumulating these correlations (using one of the techniques suggested previously) so that a single correlation represents this study ensures equal consideration will be given to another study with one sex group and one attitude measure.

Samples as Units

Using independent samples as units permits a single study to contribute more than one result if the tests are carried out on separate samples of people. For example, synthesists could consider statistical tests on males and females within the same study of rape attitudes as independent but not consider as independent two tests that used different measures of the same attitude construct given to the same people.

Using samples as independent units assumes that the largest portion of the variance shared by results in the same study comes from data collected on the same participants. This shared variance is removed (by combining results from different measures within samples) but other sources of dependency (e.g., researchers, settings) that exist at the study level are ignored. If you expect that the study context may have a large effect on study outcomes it is best to average sample sizes within studies before combining them (Borenstein, Hedges, Higgins, & Rothstein, 2009). This is because the contribution of the study to estimates of the variance in effect sizes will differ depending on whether samples or studies are used as the unit of analysis. In Chapter 6 you will learn about fixed-effect models of error (these do not vary regardless of the unit of analysis) and random-effects models for error (these do).

It is also the case that combining results based on subsamples in one study but whole samples in another can be problematic. For example, if a study of homework provides separate results for fourth and fifth grades, the average effect of homework across the two subsamples might be different from the single effect you might have obtained if the study presented one overall result. The overall effect in the study can be obtained if the group means, standard deviations, and sample sizes are available (Borenstein et al., 2009). If you have these, you can calculate them using the *Practical Meta-Analysis Effect Size Calculator* (Wilson, 2015).

When meta-analysts calculate an average comparison or relationship across units, *it is good practice to weight each independent unit—be it a sample within a study or the entire study—by its sample size.* Then, weightings are functionally equivalent whether independent samples within studies or entire studies are used as units of analysis. For example, a study with 100 participants would be weighted by 100 if the study is used as the unit, or its two samples would each be weighted by 50 if the sample is used as the unit (more will be said about this procedure in Chapter 6).

Comparisons or Estimates as Units

The least conservative approach to identifying independent units of analysis is to use each individual group comparison or estimate of relationship strength as if it were independent. That is, each separate comparison or estimate calculated by primary researchers is regarded as an independent estimate by the research synthesist. This technique's strength is that it does not lose any of the within-study information regarding potential moderators of the studies' outcomes. Its weakness is that it is likely to violate the assumption made in the meta-analytic statistical procedures that the estimates are independent. Also, the results of studies will not be weighted equally in any overall conclusion about results. Instead, studies will contribute to the overall finding in relation to the number of statistical tests contained in them, regardless of their sample size. In the example concerning rape attitudes and the need for power, the study with four related comparisons (for two sexes on two measures) will have four times the influence on the overall results as a second study with one comparison (but an equal total sample size). This is generally not a good weighting criterion.

Shifting Unit of Analysis

A compromise approach to identifying comparisons is to employ a shifting unit of analysis. Here, each outcome is initially coded as if it were an independent event. Thus, the single study that contained four estimates of the relationship between attitudes toward rape and the need for power would have four outcome coding sheets filled out for its four results. Two of these outcome code sheets (the two measures) would be linked to two different sample code sheets (the two sexes) associated with this study. Then, when an overall cumulative result for the synthesis is calculated—that is, when the question, "What is the overall relationship between attitude about rape and the need for power?" is answered—the outcome results would first be combined so that each study (requiring that all four results be combined) or each sample (combining the two outcomes for each sample) contributed equally to the overall finding. Of course, each result should still be weighted by its sample size. These combinations would then be added into the analysis across all studies.

However, the shifting unit approach allows that when examining potential moderators of the overall outcome, a study's or sample's results would be aggregated only *within* the separate categories of the moderator variable. An example should make this clearer. Suppose you have chosen to use studies as the basic unit of analysis. If a rape-attitude and need-for-power study presented correlations for males and females separately, this study would contribute only one correlation to the overall analysis—the average of the male and female correlations—but two correlations to the analysis of the impact of sex on the size of the correlation—one for the female group and one for the male group. To take the process one step farther, assume this study reported different correlations between rape attitudes and need for power within each sex for two different attitude measures—that is, four correlations in all. Then, the two correlations for different attitude scales would be averaged for each sex when the analysis examining the moderating influence of sex was conducted. Likewise, the two sex-related correlations would be averaged for each scale when the type of attitude measure was examined as a moderator.

In effect, the shifting-unit technique ensures that for analyses of influences on study estimates of relationship strength, a single study can contribute one data point to each of the categories distinguished by the moderating variable. This strategy is a good compromise that allows studies to retain their maximum information value while keeping to a minimum any violation of the assumption of independence of statistical tests. However, the approach is not without problems. First, creating and recreating average effect sizes for analysis of each different moderator can be time consuming and difficult in some statistical packages. Also, when the meta-analysts wish to study multiple influences on study outcomes in a single analysis, rather than one influence at a time, the unit of analysis can quickly decompose into individual comparisons.

The synthesis of studies examining correlates of rape attitudes included 65 research reports containing 72 studies with data on 103 independent samples. Primary researchers calculated a total of 479 correlations. Clearly, using the individual correlations as if they were independent results would grossly exaggerate their cumulative information value. For the overall analysis, then, the 103 independent samples were used as the unit and all correlations were averaged within samples. However, an analysis of differences in average correlations for different rape attitude scales was based on 108 correlations, because five primary researchers had given two scales to the same sample of participants.

Statistical Adjustment

Gleser and Olkin (2009) discuss statistical solutions to the problem of nonindependent tests. They propose several procedures that statistically adjust for interdependence among multiple outcomes within studies and for different numbers of outcomes across studies. The key to successfully using these techniques lies in the synthesists having credible estimates of the interdependence of the statistical tests. For instance, assume a study of correlates of rape attitudes includes both a measure of myth acceptance and victim blame. In order to use the statistical techniques, the synthesists must estimate the correlation between the two scales for the sample in this study. Data of this sort often are not provided by primary researchers. When not given, it might be estimated from other studies or the analysis could be run with low and high estimates to generate a range of values.

THE EFFECTS OF DATA GATHERING ON SYNTHESIS OUTCOMES

Variation in the procedures used by research synthesists to gather information from studies can lead to systematic differences in how studies are represented in the research synthesis data set. This in turn can lead to differences in what the synthesists conclude about the literature. Variation can happen in at least three ways.

First, if the synthesists only cursorily detail study operations, their conclusions may overlook important distinctions in results. A conclusion that the synthesis results indicate no important influences on study outcomes can occur either because no such influences truly exist or because the synthesists missed representing important influences in their data set. A lack of overlap in the study details considered relevant by different synthesists studying the same problem will create variation in their conclusions. However, the notion that a synthesis leads to more trustworthy results if it includes more tests of potential influences on the overall synthesis result must be tempered by the fact that the more influences tested, the more likely it is that chance alone will lead to significant findings. So, *best practice suggests that you be judicious in your choices of what influences to test*. Still, as noted before, the coding guide should be constructed to be exhaustive; not everything coded needs to be tested.

Second, synthesists can come to different conclusions about a research literature because they code studies with different accuracy. If two syntheses vary in how carefully variables are defined and coders are trained, they likely will also vary in the number of errors in their data sets, and possibly in their conclusions because of these errors. Clearly, all else being equal, the synthesis with the more rigorous coding procedures is the one with more credibility.

And finally, the conclusions of syntheses can vary because the synthesists have used different rules for judging study results as independent tests of the problem. Here, some synthesists may place greater importance on ensuring independence while others consider it more valuable to extract as much information as possible from their data.

EXERCISES

For studies on a topic of interest to you:

1. Draw up a preliminary coding guide.
2. Find several reports that describe research that is relevant to the topic.
3. Apply the coding guide to several studies, some of which you have not read before.