Meta-Analysis Notes

Jamie DeCoster

Institute for Social Science Research
University of Alabama
Box 870216
Tuscaloosa, AL 35487-0216

Phone: (205) 348-4431
Fax: (205) 348-2849

July 31, 2009

These were compiled by Jamie DeCoster, partially from a course in meta-analysis taught by Alice Eagly at Northwestern University. If you wish to cite the contents of this document, the APA reference for them would be

DeCoster, J. (2009). *Meta-Analysis Notes.* Retrieved <month, day, and year you downloaded this file> from http://www.stat-help.com/notes.html

For future versions of these notes or for help with data analysis visit
`http://www.stat-help.com`

Contents

Chapter 1

Introduction and Overview

1.1 Basics

- Definition of meta-analysis (from Glass, 1976): *The statistical analysis of a large collection of analysis results for the purpose of integrating the findings.*

- The basic purpose of meta-analysis is to provide the same methodological rigor to a literature review that we require from experimental research.

- We refer to the direct investigation of human or animal data as "primary research." A summary of primary research using statistical methodology and analysis is called "quantitative synthesis" or "meta-analysis." A summary of primary research using traditional, literary methods is called a "narrative review."

- Meta-analyses are generally centered on the relation between one explanatory and one response variable. This relation, "the effect of X on Y," defines the analysis.

- Meta-analysis provides an opportunity for *shared subjectivity* in reviews, rather than true objectivity. Authors of meta-analyses must sometimes make decisions based on their own judgment, such as when defining the boundaries of the analysis or deciding exactly how to code moderator variables. However, meta-analysis requires that these decisions are made public so they are open to criticism from other scholars.

1.2 Criticisms of Narrative Reviews

- The sample of studies examined in a narrative review is based on the author's whim, rather than on publicly shared standards.

- Narrative reviews rely on statistical significance for evaluating and comparing studies. Significance is dependent on sample size, so a weak effect can be made to look more important simply by adding more participants.

- Narrative reviews lack systematic rules regarding how to generalize from the results of individual studies to form conclusions about the literature as a whole.

- Narrative reviews are not well-suited for analyzing the impact of moderating variables. Authors of narrative reviews rarely reach clear conclusions regarding how methodological variations influence the strength of an effect. They also typically fail to report the rules they use to classify studies when looking for the effect of a moderating variable.

- Many research literatures have grown too large for a human to accurately synthesize without the aid of statistical inference.

1.3 Types of meta-analyses

- By far the most common use of meta-analysis has been in *quantitative literature reviews.* These are review articles where the authors select a research finding or "effect" that has been investigated in primary research under a large number of different circumstances. They then use meta-analysis to help them describe the overall strength of the effect, and under what circumstances it is stronger and weaker.

- Recently, as knowledge of meta-analytic techniques has become more widespread, researchers have begun to use *meta-analytic summaries* within primary research papers. In this case, meta-analysis is used to provide information supporting a specific theoretical statement, usually about the overall strength or consistency of a relation within the studies being conducted. As might be expected, calculating a meta-analytic summary is typically a much simpler procedure than performing a full quantitative literature review.

1.4 Steps to Perform a Meta-Analysis (from DeCoster, 2005)

1. Define the theoretical relation of interest.

2. Collect the population of studies that provide data on the relation.

3. Code the studies and compute effect sizes.

4. Examine the distribution of effect sizes and analyze the impact of moderating variables.

5. Interpret and report the results.

1.5 Criticisms of Meta-Analyses (and Responses)

- *Meta-analysis adds together apples and oranges.* The purpose of a literature review is to generalize over the differences in primary research. Overgeneralization can occur just as easily in narrative reviews as it can in meta-analysis.

- *Meta-analysis ignores qualitative differences between studies.* Meta-analysis does not ignore these differences, but rather codes them as moderating variables. That way their influence can be empirically tested.

- *Meta-analysis is a garbage-in, garbage-out procedure.* This is true. However, since the specific content of meta-analyses is always presented, it should be easier to detect poor meta-analyses than it would be to detect poor narrative reviews.

- *Meta-analysis ignores study quality.* The effect of study quality is typically coded as a moderator, so we can see if there is any difference between good and bad studies. If a difference does exist, low quality studies can be removed from analysis.

- *Meta-analysis cannot draw valid conclusions because only significant findings are published.* Meta-analyses are actually less affected by this bias than narrative reviews, since a good meta-analysis actively seeks unpublished findings. Narrative reviews are rarely based on an exhaustive search of the literature.

- *Meta-analysis only deals with main effects.* The effect of interactions are examined through moderator analyses.

- *Meta-analysis is regarded as objective by its proponents but really is subjective.* Meta-analysis relies on shared subjectivity rather than objectivity. While every analysis requires certain subjective decisions, these are always stated explicitly so that they are open to criticism.

Chapter 2

Formulating a Research Problem

2.1 Defining the Research Question

- There are several things you should consider when selecting a hypothesis for meta-analysis.

  1. There should be a significant available literature, and it should be in a quantifiable form.
  2. The hypothesis should not require the analysis of an overwhelming number of studies.
  3. The topic should be interesting to others.
  4. There should be some specific knowledge to be gained from the analysis. Some reasons to perform meta-analyses are to
     - Establish the presence of an effect.
     - Determine the magnitude of an effect.
     - Resolve differences in a literature.
     - Determine important moderators of an effect.

- When performing a meta-analytic summary, you often limit your interest to establishing the presence of an effect and estimating its size. However, quantitative literature reviews should generally go beyond this and determine what study characteristics moderate the strength of the effect.

- The first step to defining your research question is to decide what theoretical constructs you will use as the explanatory and response variables in your effect.

- You need to decide what effect size you will use. If the explanatory variable is typically presented as a categorical variable, you should probably use g. If the explanatory variable is typically presented as a continuous variable, you should probably use r.

- If you decide to use the effect size g, you need to precisely define what contrast you will form the basis of your effect size. For a simple design, this will probably be (experimental group - control group). Defining the contrast also specifies the directionality of your effect size (i.e., the meaning of the sign).

- If you decide to use the effect size r, you need to define the directionality of the variables to be correlated. Sometimes bipolar constructs are measured in different ways in different studies. For example, one study could use a measure of extraversion whereas another could use a measure of introversion. Both of these are measuring the same bipolar construct, but have opposite meanings. Once you specify the directionality of the variables composing your correlation, the interpretation of the sign on the correlation is automatically defined.

2.2 Limiting the Phenomenon of Interest

- Once you have determined what effect you want to examine, you must determine the population in which you want to examine it. If you are performing a meta-analytic summary you will often chose very practical boundaries for your population, such as the experiments reported in a specific paper. The populations for quantitative literature reviews, however, should be defined on a more abstract, theoretical level. In the latter case you establish a specific set of inclusion and exclusion criteria that studies must meet to be included in the analysis.

- The goal of this stage is to define a population that is a reasonable target for synthesis. You want your limits narrow enough so that the included studies are all examining the same basic phenomenon, but broad enough so that there is something to be gained by the synthesis that could not easily be obtained by looking at an individual study.

- The first criterion you must have is that the studies need to measure both the explanatory and response variables defining your effect and provide an estimate of their relation. Without this information there is nothing you can do with a study meta-analytically.

- You will also likely want to exclude studies that have not been formally written up. It can be very difficult to appropriately code a study if the details are not presented in a paper. It is also unlikely that the studies for which you can obtain the data without a writeup are a true random sample of all of the unpublished studies - typically this data will be most commonly available from researchers that you know personally. This could bias the results of your analysis.

- Each additional criterion that you use to define the population of your meta-analysis should be written down. Where possible, you should provide examples of studies that are included or excluded by the criterion to help clarify the rule.

- You should expect that your list of inclusion and exclusion criteria will change during the course of your analysis. Your perception of the literature will be better informed as you become more involved in the synthesis, and you may discover that your initial criteria either cut out parts of the literature that you want to include, or else are not strict enough to exclude certain studies that you think are fundamentally different from those you wish to analyze. You should feel free to revise your criteria whenever you feel it is necessary, but if you do so after you've started coding you must remember to recheck studies you've already completed.

Chapter 3

Searching the Literature

3.1 Basic Search Strategy

- Once you determine the boundaries of your meta-analysis, you need to locate all of the studies that fit within those bounds. When performing a meta-analytic summary you will sometimes know at the start exactly what studies you want to include. For other summaries, and for all quantitative literature reviews, you will need to perform a detailed search to locate all the studies that have examined the effect of interest within the population you defined.

- The steps to a comprehensive literature search are:

  1. Search the literature to find possible candidates for the analysis using fairly open guidelines. You should try to locate all of the studies that truly meet your criteria, even if your searches also include a large number of irrelevant studies. More specific detail on this will be provided in section 3.2.
  2. Compile a *full candidate list.* Many studies will turn up in several of your searches, so you need to combine the results into a list where each study only appears once. Reference software such as EndNote, ProCite, or Reference Manager can be helpful here, since these will allow you to automatically discard duplicate studies when combining the results from multiple search engines.
  3. Examine the title and abstract of each study in the master candidate list. Exclude any studies that are clearly not relevant to your meta-analysis. If you are uncertain as to whether a study meets your inclusion criteria based on the title and abstract, do not exclude it. The studies that make it through this initial examination are your *reduced candidate list.*
  4. Examine an electronic or paper copy of each study on the reduced candidate list to determine whether they meet your criteria for inclusion in the meta-analysis. You should start by reading the title and abstract and then continue to the methods and results sections if you need more information to make your decision. Studies that make it through this last pass are your *final candidate list.*

- You want to make sure that your full candidate list includes all of the studies you might be interested in, even if this also means including many studies that you do not use. It is not uncommon to discard over 90% of the studies from the initial list.

- The reduced candidate list should be sorted based on the source (e.g., journal or book title) when determining whether they are to be included in the meta-analysis. This way you can examine all of the studies coming from the same source at the same time, cutting out some redundant steps.

- You will need to use different methods to obtain studies found on your reduced candidate list. Some will be available electronically, some of the studies will be available at your library, some will have to be obtained through interlibrary loan, and some will have to be directly requested from the authors. Often times universities will charge a fee to provide you with a copy of a dissertation. To avoid this, you can try contacting the author to see if they will provide you with a copy of the document.

- You do not need to save copies of the studies on the full candidate list or the reduced candidate list, but you should acquire an electronic or paper copy of each study on the final candidate list. Electronic copies are preferable when they are available.

- Performing a comprehensive search of the literature involves working with a huge amount of information. You would be well-advised to make use of a spreadsheet or a database program to assist you in this task. For each study in the reduced candidate list you should record

  1. A terse reference to the study (such as journal name, volume number, and starting page number)
  2. The journal or book call number (if your library organizes its material by call number)
  3. Where you can find the study or its current retrieval status (requested from author, requested through interlibrary loan, etc.)
  4. Whether the study was included or excluded from the analysis
  5. What criteria were used as a basis for exclusion (if the study was excluded from the meta-analysis)

  It is usually not useful to create a database for the full candidate list. Studies that don't make it through the initial pass have very little chance of ever being included in the study, and so it would be a waste of your time to provide detailed documentation on them. All that you need is some documentation indicating which studies were included or excluded during the first pass.

- If you want to provide an accurate estimate of an effect it is important to find unpublished articles for your analysis. Many studies have shown that published articles typically favor significant findings over nonsignificant findings, which biases the findings of analyses based solely on published studies.

- You should include foreign studies in your analysis unless you expect that cross-cultural differences would affect the results and you lack enough foreign studies to test this difference. The Babelfish Translation website (http://babelfish.yahoo.com/) can be useful when trying to read foreign documents.

- Sometimes the number of studies that fit inside your boundaries is too large for you to analyze them all. In this case you should still perform an exhaustive search of the literature. Afterwards, you choose a random sample of the studies you found for coding and analysis.

3.2 Specific Search Procedures

- *Computerized Indices.* A number of databases are available on CD-ROM or over the internet. These will allow you to use keywords to locate articles relevant to your analysis.

  - Selecting the keywords for your search is very important. First, you should determine the basic structure of what you want in your search. For example, lets say you want to find studies that pair the terms related to "priming" with terms related to "impression formation."
  - You should next determine the synonyms that would be used for these terms in the database. For example, some researchers refer to priming effects as implicit memory effects. Similarly, researchers sometimes refer to an impression formation task as a person judgment task. You therefore may want your search to retrieve studies that use pair either "priming" or "impression formation" with either "impression formation" or "person judgment." Many indices, such as PsycInfo, publish a thesaurus that should make finding synonyms easier. If the index has predefined subject terms you should make sure that your list of synonyms includes all the relevant subject words.
  - Most indices support the use of wildcards, which you should use liberally. To locate research on priming in PsycInfo we might use the search term PRIM*, which would find studies that use the terms PRIMING, PRIMES, PRIMED, and other words beginning with PRIM.
  - You should then enter your search into the database. Each construct will be represented by a list of synonyms connected by ORs. The constructs themselves will be connected by ANDs. In the example above we might try (prim* OR implicit memory) AND (impression formation OR person judgment).

  - Be sure to use parentheses to make sure that the computer is linking your terms the way you want. For example, searching for (A OR B) AND C will give very different results from A OR (B AND C).
  - If your initial search produces a large number of irrelevant studies related to a single topic, you might try to keep them out of further searches by introducing a NOT term to your search. This will exclude all records that have the specified term in the document. For example, if our priming search produced a large number of irrelevant studies related to advertising that we wanted to exclude, we might revise our search to be (prim* OR implicit memory) AND (impression formation OR person judgment) NOT (ads OR advertising).
  - Some search engines will automatically change your search terms according to pre-specified rules. For example, if you search for a quoted term in PubMed, it will automatically remove the quotes if it doesn't find any examples of the full term in its database. Other databases will automatically change searches looking for a quoted phrase with three or more words to searches that simply have the words nearby each other. If you are conducting a search and a database gives you many more hits than you were expecting, you should check the rules that the database follows to see if it changed your search in an inappropriate way.
  - Whenever you conduct a computerized search you should record the name of the database, the years covered by the database at the time of the search, and the search terms you used. You will need to report all of this in your article.
  - The databases most commonly used by psychologists are:
    1. PsycInfo
    2. ERIC (Educational Resources Information Center)
    3. Dissertation Abstracts Online
    4. ABI/Inform (a worldwide business management and finance database)
    5. Sociological Abstracts (sociology literature)
    6. PubMed/MEDLINE (biomedical literature including health care, clinical psychology, gerontology, etc.)
    7. Mental Health Abstracts

    There are also a number of databases available within more specialized research areas.
  - You should search every computerized index that might possibly have studies related to your topic. Don't be afraid to look outside your own field. However, you should keep in mind that different indices use different terms, so you may have to define your search differently when working with different databases.

- *Descendant search.* If you can locate a small number of important studies that were performed at early dates, you can use the SSCI (Social Science Citation Index) or SCI (Science Citation Index) to locate later articles that cite them in their references. This is a nice complement to the standard search of computerized indices.

- *Ancestor search.* You should always examine the references of articles that you decide to include in your analysis to see if they contain any relevant studies of which you are unaware.

- *Research registers.* Research registers are actively maintained lists of studies centered around a common theme. Currently there are very few research registers available for psychological research, but this may change with the spread of technology.

- *Reference lists of review articles.* Previous reviews, whether they included a meta-analysis or not, are often a fruitful place to look for relevant studies.

- *Hand search of important journals.* If you find that many of your articles are coming from a specific journal, then you should go back and read through the table of contents of that journal for all of the years that there was active research on your topic. You might make use of *Current Contents*, a journal containing a listing of the table of contents of other journals.

- *Programs from professional meetings.* This is a particularly good way to locate unpublished articles, since papers presented at conferences are typically subject to a less restrictive review (and are therefore less biased towards significant findings) than journal articles. Probably the two most important conferences in psychology are the annual meetings of APA (American Psychological Association) and APS (American Psychological Society).

- *Letters to active researchers.* It is a good policy to write to the first author of each article that you decide to include in your analysis to see if they have any unpublished research relating to your topic. When trying to locate people you may want to make use of:

  - Academic department offices/Department web pages
  - Alumni offices (to track down the authors of dissertations)
  - Internet search engines (www.switchboard.com, people.yahoo.com)
  - APA or APS membership lists

Chapter 4

Coding Studies

4.1 How to Code

- Once you have collected your sample of studies you need to code their characteristics as moderator variables and calculate effect sizes.

- The steps of a good coding procedure are

  1. Decide which characteristics you want to code.
  2. Decide exactly how you will measure each characteristic. If you decide to use a continuous scale, specify the units. If you decide to use categories, specify what groups you will use.
  3. Write down the specifics of your coding scheme in a code book. The code book should contain explicit instructions on how to code each characteristic, including specific examples where necessary.
  4. Pilot the coding scheme and train the coders. You should probably code 2-4 studies between training sessions.
  5. Next you have the coders code the studies. The coders should work independently, with only occasional meetings to correct ambiguities in the scheme.
  6. Calculate the reliability of the coding for each item in your scheme. You should not include the studies you used for training in your calculation of reliability.

- Just as you might expect that your inclusion/exclusion criteria may evolve as you perform your literature search, you may also expect that your coding scheme may change as you code your studies. One important difference between primary research and meta-analysis is that your subjects (the articles) are always available and will not be influenced by repeated examination. This gives you much more flexibility in your data collection, including the opportunity for multiple reassessments, that you do not have in primary research.

- You should always have a second coder when performing a meta-analysis. Not only does this let you report a reliability on your coding of moderators, but it also provides a check on your coding and effect size calculations.

- You should prefer "low-inference" codes, where the codes are based on information that is directly reported in the study, to "high-inference" codes, where the coder has to evaluate or rate the study along a dimension. High-inference codes will typically have substantially lower reliabilities. Often times a high-inference code can be broken down into a number of low-inference codes.

- Sometimes the information you need will not be reported in a study. You should therefore have a value to indicate that the information for a particular question was unavailable. You can try contacting the authors for the information, but this often fails to gain you anything.

- Coding differences are often caused by ambiguities in the coding scheme. You should therefore concentrate on developing clear and detailed coding rules when piloting your scheme.
- Reliability is a measure of the consistency of your coding scheme. If your coding has low reliability, then the specific scheme you are using is adding a lot of variability to your measurements. It is actually a specific mathematical concept, namely

  $$\frac{\text{variability of idealized ``true'' scores}}{\text{variability of measured scores}}. \tag{4.1}$$

  Since the variability of measured scores = (variability of true scores) + (measurement error), reliabilities will always be between 0 and 1. When reporting the reliability of your coding, you should use a statistic that conforms to this definition. Some examples are (for continuous variables) the intraclass correlation, Cronbach's alpha, and (for categorical variables) Cohen's kappa.
- What is an acceptable code will vary depending on what is being coded. Ideally you would like to have all of your reliabilities at .80 or higher, but this may not be possible. You may need to explain or justify the coding scheme for any moderators that have reliabilities below .70.
- If a moderator has poor reliability, it is reasonable to have the coders meet to discuss what they did, revise the coding scheme, and then recode the studies. As long as the coders do not talk about specific studies outside of those used for training during this meaning, this second coding should not unduly bias the results.
- A computerized database can assist coding in many ways. Not only can you store the information in the database, but you can also create forms to assist in data entry, use mail-merge documents for contacting authors, print annotated copies of the data for reference, and generate output files for use by analysis programs.

4.2 What to Code

- *Study ID.* You should assign a unique number to every study included in your analysis. You should write this number on the photocopy of the study, as well as any coding or calculation sheets.
- *Long and short references.* You should record the full (APA style) reference, as well as a short citation to use when referring to the study in your notes.
- *Effect size.* This should be reported in the metric that you will be using for your analyses, even if you originally had to calculate it using a different effect size.
- *Sample sizes.* If you are working with correlations then you only need to report the overall sample size. If you are working with mean differences, you should record the sample sizes of the two groups individually.
- *Moderator variables.* You should record the codes for all of your moderator variables. Section 4.3 provides a detailed discussion of the different types of moderators you might wish to consider.
- *Characteristics of study quality.* You can then use these either as moderating variables or as bases for exclusion. One good way to code quality is to read through a list of validity threats (such as from Cook & Campbell, 1979) and consider whether each might have influenced studies in your analysis.
- *Calculation issues.* You will of course want to make detailed notes about how you calculated the effect size. More information about this will be presented in section 7.3. In addition, you may want to create codes to indicate any time you had to calculate an effect size from imperfect information. Some examples are:
  - Assuming an effect size is 0 because it was reported as nonsignificant and no other information was available.

  - Calculating an effect size from a p-value instead of a more exact statistic.
  - Estimating means from a graph.

  You may want to determine if the results that exclude these more questionable calculations are different from the results that include them. If they do not notably differ then you will want to leave the questionable calculations in the analyses. If they do, then you will have to examine the differences more closely to see which is the more appropriate set of analyses to report.

4.3 Selecting Moderators

- Sometimes there are differences between the studies that you wish to examine in your synthesis. If you code important study characteristics, you can examine whether the strength of your effect is influenced by these variables. This is called a moderator analysis.

- There are primarily three different types of moderators you will want to code in a meta-analysis.

  1. *Major methodological variations.* Your basic effect might have been examined using different procedures, different manipulations, or different response measures. The effects found with some methodologies may be stronger than with others.
  2. *Theoretical constructs.* Most literatures will come with theories that state whether the effect should be strong or weak under certain conditions. In order to address the ability of these theories to explain the results found in the literature, it is necessary to code each of your studies on theoretically important variables.
  3. *Basic study characteristics.* There are a number of moderator variables that are typically coded in any meta-analysis. These include measures of study quality, characteristics of the authors, characteristics of the research participants, and the year of publication. Generally you don't expect these variables to influence the strength of your effect, but you should always check them to rule out the possibility of them being confounding variables.

- The power of moderator analysis depends on the distribution of that variable in your sample. You will have more power when you have even numbers of studies in each level, and less when the numbers are unbalanced. If almost all of your studies have the same value on a moderator, then a test on that variable will not likely be informative. You should therefore try to select moderators that possess variability across your sample of studies.

- Just as the boundaries of your population may change as you work on your analysis, the variables that you decide to code as moderators may also change as you learn more about the literature.

- You should precisely specify exactly how each moderator will be coded. Sometimes the values that you assign to a moderator variable are fairly obvious, such as the year of publication. Other times, however, the assignment requires a greater degree of inference, such as when judging study quality. You should determine specific rules regarding how to code such "high-inference" moderators. If you have any high-inference codings that might be influenced by coder biases you should either come up with a set of low-inference codes that will provide the same information, or have the coding performed by individuals not working on the meta-analysis.

- You should make sure to code all the important study characteristics that you think might moderate your effect. There is a tradeoff, however, in that analyzing a large number of moderators does increase the chance of you finding significant findings where they don't actually exist. Statistically this is referred to as "inflating your probability of an α error." Most meta-analysts feel that it is better to code too many moderators than to code too few. If you have many moderators you might consider performing a multiple regression analysis including all of the significant predictors of effect size (see section 9.5). The results of the multiple regression automatically takes the total number of moderators into account.

4.4 Including Multiple Cases From a Single Study

- Typically each study in your sample will contribute a single case to your meta-analytic data set. Sometimes, however, a study may examine your effect under multiple levels of your moderating variables. For example, in a meta-analysis investigating the effect of priming, you might locate a study that manipulates both gender (male vs. female) and race (black vs. white), two moderating variables of interest. If you would simply calculate an overall effect from this study you would be averaging over the different levels of your moderators, so it couldn't contribute to the analysis of those variables. To take advantage of the within-study differentiation your data set would need to have several different cases for this single study.

- The simplest method to account for within-study variability is to include one case for each combination of the levels of your moderating variables. In the example above, we would have a total of four effects (male/black, male/white, female/black, female/white). Coding several cases from a single study, however, introduces a dependence in your data that must be accounted for in your analyses.

- One way to account for this dependence is to analyze the data using a mixed-effects model, where study is included in the analysis as a random factor. More information about mixed and random effects models is provided in sections 8.6 and 9.1.

- If you have multiple cases from at least some of your studies but still choose to work with a fixed-effects model, Cooper (1989) recommends that you combine together different cases that have the same level of the moderator being examined. For example, when conducting the moderator analysis for race in the example above we would calculate one effect size from white targets and one from black targets, averaging over gender. This gives us two cases from this study, instead of the four created by crossing race with gender. Similarly, we would calculate effect sizes for male targets and female targets, averaging over race, when analyzing the influence of gender on priming. For any other moderator we would use a single case for the entire study, averaging over all of the conditions.

- For tests of interactions you should use the following guidelines to determine what effect sizes to calculate.

  - If the study manipulates both of the variables in the interaction then you would want to include cases for each cell of the interaction present in the study.
  - If the study only manipulates one of the variables in the interaction you want to include cases for each level of that moderator present in the study.
  - If the study does not manipulate either of the variables in the interaction then you would have a single case representing the whole study.

- The one disadvantage of using the Cooper (1989) method is that your different moderator analyses will not all be based on the same sample. The total number of cases and the total variability in effect sizes will vary from analysis to analysis.

- If you have multiple cases from at least some of your studies you will want to divide your coding scheme into two parts.

  - *Study sheet.* Records characteristics that are always the same for cases drawn from the same study. This will include reference information and basic study characteristics.
  - *Case sheet.* Records characteristics that might be different in subsamples of a study. This will include manipulated variables and effect size characteristics.

  Each article will have a single study sheet, but may have several case sheets. Separating your coding scheme this way prevents you from recording redundant information.

- In addition to the moderating variables, your case sheet should record

  - *Case number.* Each case from a study should be given a unique identification number. References to an case would be a combination of the study number and case number ("Case 17-1").

  - *Case source.* A description what groups and responses are included in the case.
  - *Analysis inclusion codes.* For each analysis you want to perform you will need an inclusion code variable. This includes both moderator analyses as well as tests of multiple regression models. An inclusion code variable should have a value of "1" if the given case is one that should be included in the corresponding analysis. It should have a value of "0" otherwise. Having these variables will make it much easier to select the appropriate cases for each analysis.

- If you code multiple cases from each study you should consider storing your information in a relational database. A relational databases has multiple tables of information linked together by the values of specific fields. You would create separate tables to hold information from your study sheets and case sheets, and then use a "study number" field to link the two together. Using a relational database makes creating data files for your analyses much easier.

Chapter 5

Calculating Correlation Effect Sizes

5.1 Introduction

- Correlations are widely used outside of meta-analysis as a measure of the linear relation between two continuous variables. The Pearson correlation between two variables x and y may be calculated as

$$r_{xy} = \frac{\sum z_{xi} z_{yi}}{n}, \tag{5.1}$$

  where z_{xi} and z_{yi} are the standardized scores of the x and y variables for case i.

- Correlations can range between -1 and 1. Correlations near -1 indicate a strong negative relation, correlations near 1 indicate a strong positive relation, while correlations near 0 indicate no linear relation.

- The correlation coefficient r is a slightly biased estimator of ρ, the population correlation coefficient. An unbiased approximation of the population correlation coefficient may be obtained from the formula

$$G_{(r)} = r + \frac{r(1-r^2)}{2(n-3)}. \tag{5.2}$$

- The sampling distribution of a correlation coefficient is somewhat skewed, especially if the population correlation is large. It is therefore conventional in meta-analysis to convert correlations to z scores using Fisher's r-to-z transformation

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \tag{5.3}$$

  where $\ln(x)$ is the natural logarithm function. All meta-analytic calculations are then based on z_r.

- If you wish to work with unbiased estimates of ρ, you should first calculate the correction $G_{(r)}$ for each study and then transform the $G_{(r)}$ values into z_r scores for analysis using equation 5.3.

- z_r has a nearly normal distribution with variance

$${s_z}^2 = \frac{1}{n-3}. \tag{5.4}$$

- Using these statistics we can construct a level C confidence interval for the population value

$$z_r \pm \frac{z^*}{\sqrt{n-3}}, \tag{5.5}$$

  where z^* is the critical value from the normal distribution such that the area between $-z^*$ and z^* is equal to C.

- After performing a meta-analysis using z_r scores, researchers will often transform the results (such as the weighted mean effect size and confidence interval boundaries) back to the original correlation metric to make them easier to interpret. You can do this using Fisher's z-to-r transformation

$$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}, \tag{5.6}$$

  where e is the base of the natural logarithm (approximately 2.71828).

- Meta-analysts have developed formulas to calculate r from a number of different test statistics which we will present below. If you chose to use one of these formulas you should remember to correct the correlation for its sample size bias using formula 5.2, and then convert this to a z_r score using formula 5.3 before analyzing the effect sizes.

5.2 Calculating r from Linear Regression

- If a study reports the results of a simple linear regression

$$y = b_0 + b_1 x_1, \tag{5.7}$$

  you can calculate $r_{y,x1}$ using the equation

$$r_{y,x1} = b_1 \left(\frac{s_{x1}}{s_y} \right), \tag{5.8}$$

  where s_{x1} and s_y are the standard deviations of the x_1 and y variables, respectively.

  The correlation can also be obtained from the r^2 of the regression model. The correlation between x_1 and y is simply the square root of the model r^2.

- Sometimes you have the results of a multiple regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n, \tag{5.9}$$

  where your variables of interest are y and x_1. It is more difficult to calculate r in this case because the value of b_1 is affected by the other variables in the model. You can, however, use the "tracing method" to calculate $r_{y,x1}$ if you know the correlations between the predictor variables. This method is detailed in many books on structural equation modelling, including Kenny (1979, p. 31-33).

5.3 Calculating r from Test Statistics

- As mentioned above, the correlation coefficient is designed to measure the linear relation between two variables. However, there are several statistics that can be calculated from dichotomous variables that are related to correlation.

  - r_b: biserial r. This measures the relation between two continuous variables when one of them is artificially dichotomized. It is an acceptable estimate of the underlying correlation between the variables.
  - $r_{\cos-\pi}$: tetrachoric r. This measures the relation between two continuous variables when both of them are artificially dichotomized. It is also an acceptable estimate underlying correlation.
  - r_{pb}: point-biserial r. This measures the relation between a truly dichotomous variable and a continuous variable. It is actually a poor estimate of r, so we usually transform r_{pb} to r_b using the equation

$$r_b = \frac{r_{pb}\sqrt{n_e n_c}}{|z^*|(n_e + n_c)}, \tag{5.10}$$

    where z^* is the point on the normal distribution with a p-value of $\frac{n_e}{n_e+n_c}$.

  - r_ϕ: phi coefficient. This measures the relation between two truly dichotomous variables. This actually is an r.

- If you have a t statistic you can calculate r_{pb} using the formula

$$r_{pb} = \sqrt{\frac{t^2}{t^2 + n_e + n_c - 2}}. \tag{5.11}$$

  You can then transform r_{pb} into r_b using equation 5.10 to get an estimate of r.

- If you have a 1 df F statistic you can calculate r_{pb} using the formula

$$r_{pb} = \sqrt{\frac{F}{F + n_e + n_c - 2}}. \tag{5.12}$$

  You can then transform r_{pb} into r_b using equation 5.10 to get an estimate of r.

- If you have an F statistic with more than 1 df you will need to calculate a g statistic from a linear contrast of the group means and then transform this into an r. If there is an order to the groups you might consider a first-order polynomial contrast (Montgomery, 1997, p. 681), which will estimate the linear relation between your variables. See section 6.2 for more information about calculating g from linear contrasts.

- You can calculate r from the cell counts of a 2 x 2 contingency table. Consider the outcome table

	$X = 0$	$X = 1$
$Y = 0$	a	b
$Y = 1$	c	d

  where $a, b, c,$ and d are the cell frequencies. You can compute a tetrachoric r using the formula

$$r_{\cos-\pi} = \cos\left(\frac{180^\circ}{1 + \sqrt{\frac{ad}{bc}}}\right). \tag{5.13}$$

- If you have a 2 x 2 table for the response frequencies within two truly dichotomous variables, you can calculate r_ϕ from a chi-square test using the equation

$$r_\phi = \sqrt{\frac{\chi^2}{n}}. \tag{5.14}$$

- If you have a Mann-Whitney U (a rank-order statistic) you can calculate r_{pb} using the formula

$$r_{pb} = 1 - \frac{2U}{n_e n_c}, \tag{5.15}$$

  where n_e and n_c are the sample sizes of your two groups. To get an estimate of r you can then transform r_{pb} to r_b using equation 5.10.

5.4 Miscellaneous

- You can calculate r from g using the equation

$$r = \sqrt{\frac{g^2 n_e n_c}{g^2 n_e n_c + (n_e + n_c)(n_e + n_c - 2)}}. \tag{5.16}$$

- You can calculate r from g^* using the equation

$$r = \sqrt{\frac{g^{*2}}{g^{*2} + 4}}, \tag{5.17}$$

assuming that you have approximately the same number of subjects in the experimental and control groups. If the populations are clearly different in size, then you should use the equation

$$r = \sqrt{\frac{g^{*2}}{g^{*2} + \frac{1}{pq}}}, \tag{5.18}$$

where $p = \frac{n_e}{n_e + n_c}$ and $q = 1 - p$.

Chapter 6

Calculating Mean Difference Effect Sizes

6.1 Introduction

- In this chapter we will discuss how to calculate estimates of the standardized mean difference effect size δ. δ is meant to represent the magnitude of the difference between two populations. This effect size is calculated using the formula
$$\delta = \frac{\bar{\mu}_1 - \bar{\mu}_2}{\sigma}, \tag{6.1}$$
where μ_1 and μ_2 are the means of the two populations and σ is either the standard deviation of one of the two populations, or the pooled standard deviation of both populations.

- The effect size δ is most commonly used to represent the size of an effect observed in a t-test. For convenience sake we will assume that μ_1 represents the mean of an experimental group and μ_2 represents the mean of a control group. When considering the role of this difference in the design of the study, we will call the variable differentiating these groups as the "treatment factor."

- Several different statistics have been developed to estimate δ. The equations for the statistics all take the form of a fraction with the mean difference between the two groups in the numerator and a standard deviation in the denominator. The differences between the statistics lies in what standard deviation they use in the denominator.

  ◦ Glass's Δ is calculated using the formula
  $$\Delta = \frac{\bar{Y}_e - \bar{Y}_c}{s_c}, \tag{6.2}$$
  where $\bar{Y}_e$ is the mean of the experimental group, $\bar{Y}_c$ is the mean of the control group, and s_c is the standard deviation of the control group.

  ◦ Cohen's d is calculated using the formula
  $$d = \frac{\bar{Y}_e - \bar{Y}_c}{s_p}, \tag{6.3}$$
  where $\bar{Y}_e$ is the mean of the experimental group, $\bar{Y}_c$ is the mean of the control group, and s_p is the pooled sample standard deviation. For Cohen's d, the pooled standard deviation is calculated using the formula
  $$s_p = \sqrt{\frac{(n_e - 1){s_e}^2 + (n_c - 1){s_c}^2}{n_e + n_c}}. \tag{6.4}$$
  which is the biased maximum likelihood estimate of the pooled standard deviation.

- Hedges' g is calculated using the formula

  $$g = \frac{\bar{Y}_e - \bar{Y}_c}{s_p}, \tag{6.5}$$

  where $\bar{Y}_e$ is the mean of the experimental group, $\bar{Y}_c$ is the mean of the control group, and s_p is the pooled sample standard deviation. You will notice that this is the same formula that is used to calculate Cohen's d. However, Hedges' g uses calculates the pooled standard deviation using the formula

  $$s_p = \sqrt{\frac{(n_e - 1){s_e}^2 + (n_c - 1){s_c}^2}{n_e + n_c - 2}}. \tag{6.6}$$

  which is the unbiased least squares estimate of the pooled standard deviation.

- All three statistics are comparable to each other, and will be approximately equal in large samples. In this section we will focus on using Hedges' g to estimate δ.

  - Hedges' g is preferable to Glass's Δ because experimental designs make the assumption that the standard deviations of the different groups are all the same. In this case, pooling the standard deviations from both groups, as is done when calculating Hedges' g, will give us a better estimate of the population standard deviation than relying solely on the standard deviation of the control group, as is done when calculating Glass's Δ.
  - Hedges' g is preferable to Cohen's d because it uses the unbiased least squares estimate of the pooled standard deviation. Most of the analyses from which you will be deriving standardized mean difference effect sizes, such as t-tests and ANOVA, are based on the unbiased least squares estimate of the pooled standard deviation. This makes it much easier to calculate Hedges' g than Cohen's d from their statistics.

- The effect size g is actually a biased estimator of the population effect size δ. Using g produces estimates that are too large, especially with small samples. To correct g we multiply it by a correction term

  $$J_m = 1 - \frac{3}{4m - 1}, \tag{6.7}$$

  where $m = n_e + n_c - 2$. The resulting statistic

  $$g^* = g\left(1 - \frac{3}{4m - 1}\right) = g\left(1 - \frac{3}{4(n_e + n_c) - 9}\right) \tag{6.8}$$

  is an unbiased estimator of δ. It is generally best to record both g and g^* for each effect in your meta-analysis, but then base your analyses on g^*.

  The statistic $g*$ is sometimes (such as in earlier versions of these notes) referred to as "Hedges' d," but we will stick to calling it g^* to prevent confusion with Cohen's d, which is a different statistic.

- The variance of g^*, given relatively large samples, is

  $$\sigma^2_{g^*} = \frac{n_e + n_c}{n_e n_c} + \frac{{g^*}^2}{2(n_e + n_c)}. \tag{6.9}$$

- Using these statistics, we can construct a 95% confidence interval for δ using the equation

  $$g^* \pm 1.96(\sigma_{g^*}). \tag{6.10}$$

  If you want a confidence interval that is greater or less than 95%, you can just replace the 1.96 in formula 6.10 with the value z from the standard normal distribution such that the area between $-z$ and z is equal to the size of the interval you want.

- Meta-analysts have also developed formulas to calculate g from a number of different test statistics which we will present below. If you chose to use one of these formulas you should remember to correct g for its sample size bias using formula 6.8 presented above.

6.2 Calculating g from Between-Subjects Test Statistics

- If you have access to the means and standard deviations of your two groups, you can calculate g from the definitional formula
$$g = \frac{\bar{Y}_e - \bar{Y}_c}{s_p}, \tag{6.11}$$
where $\bar{Y}_e$ is the mean of the experimental group, $\bar{Y}_c$ is the mean of the control group, and s_p is the pooled standard deviation. The pooled standard deviation can be calculated from the standard deviations of your two groups using the formula
$$s_p = \sqrt{\frac{(n_e - 1){s_e}^2 + (n_c - 1){s_c}^2}{n_e + n_c - 2}}. \tag{6.12}$$
You can also use $\sqrt{\text{MSE}}$ from a one-way ANOVA model testing the treatment effect to estimate the pooled standard deviation.

- If you have a between-subjects t statistic comparing the experimental and control groups,
$$g = t\sqrt{\frac{1}{n_e} + \frac{1}{n_c}} = t\sqrt{\frac{n_e + n_c}{n_e n_c}}. \tag{6.13}$$
When you have the same number of subjects in the experimental and control group this equation resolves to
$$g = t\sqrt{\frac{2}{n}} = \frac{2t}{\sqrt{2n}}. \tag{6.14}$$

- From the same logic, if you have a between-subjects z test comparing the experimental and control groups,
$$g = z\sqrt{\frac{n_e + n_c}{n_e n_c}}. \tag{6.15}$$
When you have the same number of subjects in the experimental and control group this equation resolves to
$$g = \frac{2z}{\sqrt{2n}}. \tag{6.16}$$

- If you have a 1 numerator df F statistic comparing the experimental and control groups (we never directly calculate g from F statistics with more than 1 numerator df),
$$g = \sqrt{\frac{F(n_e + n_c)}{n_e n_c}}. \tag{6.17}$$
If you have the same number of subjects in the experimental and control groups, this equation resolves to
$$g = \sqrt{\frac{2F}{n}}. \tag{6.18}$$
Since F statistics ignore direction, these calculations will always produce positive values. You must therefore check which mean is higher and give the appropriate sign to g by hand.

  Notice the similarity between these equations and equations 6.13 and 6.14. This is because a 1 df F statistic comparing the experimental and control group will always be equal to the square of a t statistic comparing the two groups.

- If your treatment factor has more than 1 df you may choose to calculate g from a combination of the group means. In this case you

   1. Calculate the linear contrast
   $$L = \sum c_j \bar{Y}_j, \tag{6.19}$$
   where the summation is over the levels of the treatment factor, $\bar{Y}_j$ is the sample mean of group j, c_j is the coefficient for group j, and $\sum c_j = 0$.
   2. Calculate the pooled standard error
   $$s_p = \sqrt{\frac{\sum {s_j}^2 {c_j}^2 (n_j - 1)}{\sum {c_j}^2 (n_j - 1)}}, \tag{6.20}$$
   where the summation is of the levels of the treatment factor, s_j is the standard deviation of group j, and n_j is the sample size of group j.
   3. Calculate the effect size
   $$g = \frac{L}{s_p}. \tag{6.21}$$

  If you want to compare the experimental and control group, you would set the c_j for the experimental group equal to 1 and the c_j for the control group equal to -1. However, you can also use this same basic formula to calculate effect sizes for more complicated comparisons.

- Sometimes you will only know the total number of subjects run in a study, rather than how many were in each level of the design. In this case you will generally assume that there were an equal number of subjects run in each condition. This may lead to non-integer sample size estimates, but this is not a problem since the formulas will still work with these values.

6.3 Calculating g Indirectly

- Sometimes a primary research study will test a model including the treatment factor but not report a statistic specifically testing the difference between the experimental and the control group you are interested in. In this case, you can usually derive g from your comparison as long as the study reports the cell means.

- Consider a simple two-way ANOVA design:

	A_1	A_2	$\cdots$	A_a
B_1	AB_{11}	AB_{21}	$\cdots$	AB_{a1}
B_2	AB_{12}	AB_{22}	$\cdots$	AB_{a2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
B_b	AB_{1b}	AB_{2b}	$\cdots$	AB_{ab}

  From this layout we can see that factor A has a levels and factor B has b levels. To calculate a g comparing either two marginal means or two cell means in this design you can take the following steps.

   1. Determine the two means that you want to compare in your effect size. You can average together different cell means to obtain marginal means if necessary. If you cannot determine these means and you don't have a test directly comparing the means then you will not be able to calculate an effect size.
   2. Find one effect in the design for which you have both the F statistic and the means that were compared to create the F statistic. In the example above, you could have F_A and all of the marginal means for Factor A, F_B and all of the marginal means for factor B, or F_{AB} and all of the individual cell means.

3. Calculate the mean squares for the effect you chose. In a 2-way ANOVA, the mean squares for the main effects would be calculated using the formulas

$$\text{MS}_A = \frac{\sum \left[n_j \left(\bar{A}_j - \bar{G} \right)^2 \right]}{a-1}, \tag{6.22}$$

where the summation is over the different levels of A, n_j is the number of subjects in level j, and $\bar{G}$ is the grand mean, and

$$\text{MS}_B = \frac{\sum \left[n_k \left(\bar{B}_k - \bar{G} \right)^2 \right]}{b-1} \tag{6.23}$$

where the summation is over the different levels of B and n_k is the number of subjects in level k. The mean squares for the interaction effect would be calculated using the formula

$$\text{MS}_{AB} = \frac{\sum \sum \left[n_{jk} \left(\overline{AB}_{jk} - \bar{A}_j - \bar{B}_k + \bar{G} \right)^2 \right]}{(a-1)(b-1)}, \tag{6.24}$$

where the first summation is over the different levels of A and the second is over the different levels of B, and n_{jk} is the number of subjects in cell AB_{jk}.

4. Use the F statistic and the mean squares for the effect to derive the mean squared error. Recall that every F statistic in a between-subjects ANOVA can be calculated as

$$F = \frac{\text{MS}_{\text{effect}}}{\text{MS}_{\text{error}}}. \tag{6.25}$$

Using some algebra, we can use this formula to derive the equation

$$\text{MS}_{\text{error}} = \frac{\text{MS}_{\text{effect}}}{F}, \tag{6.26}$$

which we can use to calculate the mean squared error.

5. Calculate the pooled standard deviation using the equation

$$s_p = \sqrt{\text{MS}_{\text{error}}}. \tag{6.27}$$

6. At this point we have estimates of the group means and the pooled standard deviation, which we can use to calculate g from equation 6.11.

- The pooled standard deviation calculated using equation 6.27 is an estimate of the within-cell variance. In a multifactor study, however, this specifically does not include variability associated with other factors in the ANOVA. It is important to consider whether the within-cell variance actually represents the variability that you want the pooled standard deviation to capture. Sometimes it is appropriate to exclude the variability associated with the other factors, such as when the they represent manipulations that add variability to the natural situation. Other times you would actually want to include the variability associated with the other factors in your estimate of the pooled standard deviation, such as when the factors represent individual differences or other measurements of naturally varying characteristics.

- If you are in a situation where the ANOVA is removing variability that you want included in your pooled standard deviation, you can "reconstitute" the mean squared error by putting back the variance associated with other factors in the model (Johnson & Eagly, 2000). The procedure is to add the sums of squares and the degrees of freedom from the factors you want to reconstitute to those of the error term and recalculate the mean squared error. This can be represented using the formula

$$s_p^* = \sqrt{\frac{SS_1 + SS_2 + \cdots SS_k + SS_E}{df_1 + df_2 + \cdots df_k + df_E}}, \tag{6.28}$$

where s_p^* is the reconstituted pooled standard deviation, $SS_1 + SS_2 + \cdots SS_k$ are the sums of squares from the factors you want to reconstitute, and $df_1 + df_2 + \cdots df_k$ are the degrees of freedom from the factors you want to reconstitute. If you reconstitute a factor into your pooled standard deviation, you should also reconstitute all interactions involving that factor into your pooled standard deviation.

- This procedure is easy if you have a complete ANOVA table available. If you don't, you can reconstruct it yourself if you have the cell means and at least one F statistic. To recreate the ANOVA table you would take the following steps.

  1. Calculate the mean squares of your effects (such as by using using equations 6.22, 6.23, and 6.24).
  2. Determine the mean squared error from the original design using equation 6.26.
  3. Determine the degrees of freedom associated with each effect. For main effects, the df will be equal to (number of groups - 1). For interactions, the df can be calculated by multiplying together the df from all of the main effects involved in the interaction.
  4. Determine the sums of squares for each of your effects and the error term by multiplying the mean squares by the degrees of freedom.

- You have a similar issue if you are only provided with the means and standard deviations on subgroups within the experimental and control groups. Any variability associated with the dividing factor has does not affect the cell standard deviation. If you want your pooled standard deviation to reflect this variability, you can calculate s_p^* by taking the following steps.

  1. Calculate $\bar{Y}_e$ and $\bar{Y}_c$ (combining across the dividing factor) using weighted averages.
  2. Calculate the sum of squared scores for the jth subgroup within the experimental and control conditions using the equations
  $$SE_j = (n_{ej} - 1){s_{ej}}^2 + n_{ej}(\bar{Y}_{ej})^2 \tag{6.29}$$
  and
  $$SC_j = (n_{cj} - 1){s_{cj}}^2 + n_{cj}(\bar{Y}_{cj})^2, \tag{6.30}$$
  where n_{ej}, s_{ej}, and $\bar{Y}_{ej}$ are the sample size, standard deviation, and sample mean of the jth subgroup under the experimental condition, and n_{cj}, s_{cj}, and $\bar{Y}_{cj}$ are the sample size, standard deviation, and sample mean of the jth subgroup under the control condition.
  3. Add these up to get the total sum of squares within the conditions using the formulas
  $$SE = \sum SE_j - n_e(\bar{Y}_e)^2 \tag{6.31}$$
  and
  $$SC = \sum SC_j - n_c(\bar{Y}_c)^2, \tag{6.32}$$
  where n_e and n_c are the total sample sizes (aggregating across the dividing factor) within the experimental and control groups, respectively.
  4. Calculate the reconstituted pooled standard deviation using the formula
  $$s_p^* = \sqrt{\frac{SE + SC}{n_e + n_c - 2}}. \tag{6.33}$$

- If you have the means and standard deviations reported within subgroups of the experimental and control conditions and the subgroups represent a manipulated factor, it may be appropriate to calculate the effect size within each of the subgroups and then average them together, weighting the average by the sample size. This effectively removes the variability associated with the manipulated factor from the pooled standard deviation. This is appropriate when you believe the factor is adding extraneous variability to the design.

- You might encounter an ANOVA that based its analysis on difference scores (as opposed to posttest scores). If you want to calculate an effect size based on posttest scores (to make it comparable to others you calculate) you can

  1. Calculate the standard deviation of the difference scores s_{dif}.
  2. Calculate the standard deviation of the post scores using the equation
  $$s_y = \frac{s_{\text{dif}}}{\sqrt{2(1-r_{xy})}}, \tag{6.34}$$
  where r_{xy} is the correlation between the pretest and posttest scores.
  3. Calculate the effect size using the equation
  $$g = \frac{\bar{Y}_e - \bar{Y}_c}{s_y} = \frac{\overline{\text{DIF}}_e - \overline{\text{DIF}}_c}{\frac{s_{\text{dif}}}{\sqrt{2(1-r_{xy})}}}, \tag{6.35}$$
  where $\overline{\text{DIF}}_e$ and $\overline{\text{DIF}}_c$ are the mean difference scores for the experimental and control group, respectively. We get the last part of the equality from the fact that $\bar{Y}_e - \bar{Y}_c = \overline{\text{DIF}}_e - \overline{\text{DIF}}_c$.

  Note that this solution requires r_{xy}, which is not available for many studies. If the study does not report this correlation you can take it from a different study, or make a rough approximation of it to complete your calculations.

6.4 Calculating g from a within-subjects design

- The logic behind the calculation g for within-subjects comparisons is the same as that for between-subjects comparisons. However, the s used in within-subjects analyses is typically based on the standard deviation of the difference score, s_{e-c}, rather than the pooled standard deviation. The general formula for g in within-subject designs is
$$g = \frac{\bar{Y}_e - \bar{Y}_c}{s_{e-c}}. \tag{6.36}$$

- You can calculate the effect size from within-subjects test statistics using the formulas
$$g = \frac{t}{\sqrt{n}} \tag{6.37}$$
and
$$g = \frac{z}{\sqrt{n}}. \tag{6.38}$$

- You can derive an estimate of s_{e-c} from a within-subjects ANOVA table, but the procedure is a little different than with a between-subjects ANOVA. To calculate s_{e-c} you must first determine from which error mean squares it should be taken. A within-subjects ANOVA has a number of different error mean squares, and you need to choose the one that would be appropriate to test the contrast in which you are interested. The appropriate mean squares are those from the denominator of the F statistic for the factor containing the groups you are comparing. You should see a book on experimental design (such as Montgomery, 1997, or Neter, Kutner, Nachtsheim, & Wasserman, 1996) if you are not familiar with how within-subjects tests are performed.

  Once the value of this error mean squares is obtained, you can calculate s_{e-c} using the equation
  $$s_{e-c} = \sqrt{2 * \text{MS(within error)}}, \tag{6.39}$$
  where MS(within error) is the appropriate within-subjects error term.

- There is an algorithm that will tell you which effects are tested using which error mean squares in a within-subjects design. The steps to the algorithm are listed below.

    1. On the first line of a sheet of paper write down the first between-subjects factor. If there are no between-subjects factors skip to step 4.
    2. Write down another between subjects factor. Following it, write down all of the interactions between this new factor and all of the terms (including both main effects and interactions) you have already written down on the paper.
    3. Repeat step 2 until you have written down all of your between-subjects factors. At this point you should have all the between-subject main effects and interactions listed out on the top line.
    4. At the end of the same line write down "S(*interaction*)" where *interaction* is the interaction between all of your between-subjects factors. This is your between-subjects error term.
    5. On the next line, write down a within-subjects factor. Following it, write down the interaction between this new factor and every term (whether a main effect, interaction, or error term) that you have already written down on the page. At the end of the line you should have a term crossing your within-subjects factor with the between error term.
    6. On the next line write down another within-subjects factor. Following it, again write down the interaction between this new factor and every term that you have already written down on the page. This should include both the between-subjects and the within-subjects terms. Every time you write down an error term (any term that has an "S" in it) write your next term on a new line.
    7. Repeat step 6 until you have written down all of your within-subjects factors.
    8. When you are finished you should have a full list of every term in your design. Main effects and interactions that are on the same line are all tested by the same error term, which is the term listed at the end of the line.

    So that you can have an idea of how this procedure actually works, figure 6.4 contains the output when the algorithm is used on a design with 2 between-subjects factors (A and B) and 3 within-subject factors (C, D, and E). You can see that this list contains every term in the model exactly once, matched with the appropriate error term.

    A, B, A*B, S(A*B)
    C, C*A, C*B, C*A*B, C*S(A*B)
    D, D*A, D*B, D*A*B, D*S(A*B)
    D*C, D*C*A, D*C*B, D*C*A*B, D*C*S(A*B)
    E, E*A, E*B, E*A*B, E*S(A*B)
    E*C, E*C*A, E*C*B, E*C*A*B, E*C*S(A*B)
    E*D, E*D*A, E*D*B, E*D*A*B, E*D*S(A*B)
    E*D*C, E*D*C*A, E*D*C*B, E*D*C*A*B, E*D*C*S(A*B)

    Figure 6.1: Example output from the algorithm.

- Just as in between-subjects designs, you can use a different but related F statistic to indirectly calculate s_{e-c}. When performing this procedure you need to keep three things in mind.

    1. This procedure only works if the F statistic you have uses the same within error term that is appropriate for your contrast. Any other Fs will lead to incorrect estimates of s_{e-c}.
    2. Within-subjects factors have different formulas for degrees of freedom than between-subjects factors. You need to take this into consideration when calculating mean squares.
    3. Once you calculate MS(within error) you need to use formula 6.39 to get the standard deviation of the difference score.

- A within-subjects contrast calculated within the levels of a between-subjects variable uses the relevant within-subjects error term. This rule is valid even when the within-subjects contrast is calculated within crossed levels of two between variables. Therefore, the standard deviation for the denominator of g would be calculated from the within-subjects error term using equation 6.39 above.

- If you want to calculate a between-subjects contrast within the levels of a within-subjects variable, you will need to use a mixed error term. This error term would be a weighted average of the between-subjects error term and the relevant within-subjects error term. Therefore, if an effect size for a between-subjects contrast is calculated from the values for a particular level of a within-subjects variable, the standard deviation calculated for the denominator of the g would need to be an average the between-subjects pooled standard deviation and the within-subjects standard deviation of differences. If an effect size for a between-subjects contrast is calculated from the values found within a particular combination two within-subjects variables, all of the standard deviations that are derived from these error terms would be averaged (i.e., the between-subjects pooled standard deviation and the three relevant within-subjects standard deviations of differences) to create a "within-cell" standard deviation. In all cases the weighted averaging should be performed on the variances, and then the square root should be taken to produce the standard deviation for the denominator of the effect size. See Montgomery (1997) for more information about the error terms for contrasts in a mixed design.

- Effect sizes calculated from studies using a between-subjects design are not on the same metric as those calculated from studies using a within-subjects design. You should therefore never mix the two different types of effects in the same meta-analysis without compensating for these differences. A detailed discussion of these issues is provided by Morris and DeShon (2002).

  Between-subjects and within-subjects tests can differ in both the means that are being compared as well as what standard deviation is used in the tests.

  - Between-subjects tests compare the raw means of different groups of participants (e.g., $\bar{X}_e - \bar{X}_c$). Within-subjects tests will compare change scores within participants (e.g., $\bar{X}_2 - \bar{X}_1$), or will compare the change scores found within one group to the change scores found in other groups (e.g., $[\bar{X}_{e2} - \bar{X}_{e1}] - [\bar{X}_{c2} - \bar{X}_{c1}]$).
  - Between-subjects tests use the pooled standard deviation s_p, which is conceptually similar (but not mathematically equal) to an average of the individual group standard deviations. Within-subjects tests use the standard deviation of the difference score between conditions s_{e-c}.

- In order for between-subjects and within-subjects effects to be used in the same design, you must take steps to make them comparable to each other.

  - Effect sizes must be based on equivalent mean differences. In situations where there are no pretest differences between groups and no sources of change over time due to factors unrelated to the treatment, all three mean differences are equivalent. In the presence of pretest differences then the between-subjects tests will be biased while the other two will not. If there are consistent sources of change over time then within-subjects comparisons without a comparison group will be biased while the other two will not. If you can estimate the size and direction of these biases then you can correct the mean differences to make them more equivalent. Morris and DeShon (2002) review methods of meta-analytically estimating these and other biases.
  - Effect sizes must all be based on the same standard deviation for them to be comparable. The pooled standard deviation can be calculated from the standard deviation of the difference scores using the formula

    $$s_p = \frac{s_{e-c}}{\sqrt{2(1 - r_{ec})}}, \tag{6.40}$$

    where r_{ec} is the correlation between the experimental and control scores. Similarly, the standard deviation of the difference scores can be calculated using the formula

    $$s_{e-c} = s_p\sqrt{2(1 - r_{ec})}. \tag{6.41}$$

- ○ Morris and DeShon (2002) note that the variance of the estimated effect size will depend on both the study design as well as the metric that is chosen for the mean difference (i.e., raw scores or change scores). They provide formulas that will allow you to calculate these variances for different design-metric combinations. Meta-analyses that examine a combination of between-subjects and within-subjects effects should use these formulas for the variance instead of those discussed in these notes.

6.5 Calculating g from dichotomous dependent variables

- A dichotomous dependent variable is one that records whether a particular event occurs or does not occur. Some examples of dichotomous measures would be a medical study that considers whether a patient lives or dies, or a psychology study that considers whether a bystander helps or ignores a lost child.

- In this section we will discuss how to calculate g from these measures. If most of the studies in your literature use dichotomous dependent variables you should probably base your calculations on a rate-based effect size such as the odds ratio. This is covered in detail in Fleiss (1994).

- One method to calculate g from categorical data, proposed by Glass, McGaw, and Smith (1981), assumes that the dichotomous decision is based on the comparison of some underlying continuous variable (with a normal distribution) to a fixed criterion. To calculate g using this method you

  1. Choose one of the outcomes as your "critical event". This decision is arbitrary and will not affect your results.
  2. Calculate the probabilities of the critical event in your experimental group (p_e) and control group (p_c).
  3. Find the z scores z_e and z_c that correspond to these probabilities from a normal distribution table.
  4. Since the difference of z scores is also a z score, you can calculate your effect size using the equation

  $$g = (z_e - z_c)\sqrt{\frac{n_e + n_c}{n_e n_c}}. \tag{6.42}$$

  When you have the same number of subjects in the experimental and control group this equation resolves to

  $$g = \frac{2(z_e - z_c)}{\sqrt{2n}}. \tag{6.43}$$

- A second method treats the proportions of observations in each group as means of a distribution of 1's (where a critical event occurred) and 0's (where the critical event did not occur). To calculate g using this method you

  1. Choose one of the outcomes as your "critical event". This decision is arbitrary and will not affect your results.
  2. Calculate the probabilities of the critical event in your experimental group (p_e) and control group (p_c).
  3. Calculate the mean and standard deviation for each group using the equations

  $$\bar{Y} = p \tag{6.44}$$

  and

  $$s = \sqrt{pq}, \tag{6.45}$$

  where q is defined as $1 - p$.

4. Calculate the pooled standard deviation, using equation 6.12. This equation becomes

$$s_p = \sqrt{\frac{(n_e - 1)p_e q_e + (n_c - 1)p_c q_c}{n_e + n_c - 2}} \tag{6.46}$$

5. Use $\bar{Y}_e$, $\bar{Y}_c$, and s_p to calculate g using equation 6.11.

- If you do not have the actual frequencies or proportions, you can calculate an effect size from a chi-square statistic testing a difference between the two proportions.
  - If you have a 2 x 2 table, then $\chi^2 = z^2$. You may therefore get an unbiased estimate of the effect size from the equation

    $$g = \sqrt{\frac{\chi^2(n_e + n_c)}{n_e n_c}}. \tag{6.47}$$

    When you have the same number of subjects in the experimental and control group this equation resolves to

    $$g = \sqrt{\frac{2\chi^2}{n}}. \tag{6.48}$$

    You can alternatively calculate the phi coefficient using the equation

    $$r_\phi = \sqrt{\frac{\chi^2}{n}} \tag{6.49}$$

    and calculate g from r using equation 6.51.
  - If one or both variables has more than 2 levels, Glass, McGaw, and Smith (p. 150) assert that you can calculate

    $$P = \sqrt{\frac{\chi^2}{n + \chi^2}}, \tag{6.50}$$

    which approximates r if the sample size is large. If the sample size is small, then P will be too low. You can then transform r to g using equation 6.51.

6.6 Miscellaneous

- To calculate g from r you use the formula

  $$g = \frac{2r}{\sqrt{1 - r^2}}. \tag{6.51}$$

- To calculate g from nonparametric statistics you can find the p-value associated with the test and solve it for t (using the procedures discussed in section 7.1). You then calculate g using equation 6.13. For more precision you can make an adjustment for the lower power of the nonparametric statistic (see Glass, McGaw, & Smith, 1981).

Chapter 7

General Issues in Calculating Effect Sizes

7.1 Estimating effect sizes from p-values

- If you only have a p-value from a test statistic, you can calculate g if you know the direction of the finding. The basic procedure is to determine the test statistic corresponding to the p-value in a distribution table, and then calculate the effect size from the test statistic.
- You can get inverse probability distributions from a number of statistical software packages, including Excel, SPSS and SAS. Even some hand-held calculators will provide the inverse distribution of the simpler statistics.
- While an exact p-value allows an excellent estimate of a test statistic (and therefore the effect size), a significance level (e.g., $p < .05$) gives a poorer estimate. You would treat significance levels as if it were an exact p-value in your calculations (e.g., treat $p < .05$ as $p = .05$).
- The mere statement that a finding is "significant" can be treated as $p = .05$ in studies that use the conventional .05 significance level. These estimates, however, are typically quite poor.
- One problem is how do deal with a report that simply states that the effect of interest is "nonsignificant." It is common to represent such effects by setting $g = 0$ or $r = 0$, but such estimates are obviously very poor. If you have many of these reports in your data set you may want to estimate mean effect sizes with and without these zero values. This effectively sets upper and lower bounds for your mean effect size. You may want to omit these zero values when performing moderator analyses.

7.2 Choosing a Calculation Method

- There are a large number of equations to calculate effect sizes. Sometimes there is only one correct way to calculate an effect size from a given study, but other times you have a choice of several different methods. The methods, however, differ in their precision and the number of assumptions they have to make. In general, you want to calculate your effect size as directly as possible. The more inferences you have to make, the more error you will likely include in your estimate.
- *Calculating mean difference effect sizes.* The best methods calculate g from
  - $\bar{Y}_e$, $\bar{Y}_c$, and s_p for the effect of interest, where s_p is calculated by pooling
  - $\bar{Y}_e$, $\bar{Y}_c$, and s_{e-c} from a within-subjects design
  - a t test for the effect of interest
  - a 1 df F test of the effect of interest

    - proportions for experimental and control group
    - a correlation between the appropriate pair of variables

  The second class of methods calculate g from

    - $\bar{Y}_e$, $\bar{Y}_c$, and s_p, where s_p is calculated from a different but related t or g
    - $\bar{Y}_e$, $\bar{Y}_c$, and s_p, where s_p is calculated from the standard deviation of subgroups
    - $\bar{Y}_e$, $\bar{Y}_c$, and s_p, where s_p is calculated from the reconstruction of an ANOVA table based on a related F
    - a reported chi-square test

  The third class of methods calculate g from

    - a p-value
    - $\bar{Y}_e$, $\bar{Y}_c$, and s_p or s_{e-c}, where the standard deviation is calculated from the reconstruction of an ANOVA table based on a related p-value

  As a final option you can assign $g = 0$ when a study reports null effect and you can't calculate a more specific effect size.

- *Calculating correlation effect sizes.* Unlike g, r is often directly reported. The best methods are to calculate r from

    - a directly reported correlation
    - a simple linear regression coefficient

  The second class of methods calculate r from

    - a multiple regression coefficient (adjusted to no longer represent a partial effect)
    - a t test
    - a 1 df F test
    - a 2 x 2 table
    - a Mann-Whitney U statistic

  The third class of methods calculate r from

    - the dichotomization of an F statistic with more than 1 df
    - the estimation of a linear relation in an F statistic with more than 1 df
    - a p-value

  As a final option you can assign $r = 0$ when a study reports null effect and you can't calculate a more specific correlation.

7.3 Documenting Effect Size Calculations

- You should record the details of all of your effect size calculations. This will be helpful both when you are comparing effect sizes between coders and for making corrections to your calculations.

- One of the easier ways to organize your documentation is to have a single word processing document for each coder. If you create a heading for each study, you can have the word processor automatically generate a table of contents at the beginning of the document. This will enable you to quickly find whatever study you are looking for.

- The documentation for each study should contain the following items.

- The values reported in the study that you used to calculate the effect size.
- The page numbers containing the values you used to calculate the effect size.
- The overall sample size for r or the sample sizes of the experimental and control groups for g.
- Any intermediate computations you made in the process of calculating the effect size.
- Notes explaining unusual methods that you used to calculate the effect size.
- Notes explaining any difficulties you had calculating the effect size, such as missing information.
- Calculated r or g.
- If you are not using a program that automatically transforms and corrects these effect sizes, you should report z_r and g^*. There is no need to separately calculate these values if you are working with a program (such as Comprehensive Meta Analysis) that computes these automatically.
- For experimental studies, you should record information about the overall design of the study. This should be an exact specification of the study design, describing which factors are crossed and nested. Example: Time X S(Gender X Treatment). You should specify all aspects of the design, not just those relevant to your analysis.

7.4 Correcting Effect Sizes for Attenuation

- Sometimes it is useful to think of two effect size parameters, one representing the effect size found in research studies and one representing the true theoretical effect size found under ideal conditions. Our practical instantiation of research methods can never reach the ideal, so the study effect size is always somewhat less than the ideal effect size (assuming that the deviations vary randomly between studies).

- If you want to draw inferences about the theoretical effect size you need to correct your calculated effect sizes for attenuation from methodological deficiencies. For each study you must therefore calculate both a raw effect size as well as an effect size corrected for attenuation. You can then analyze the corrected effect sizes in the same way you analyze standard effect sizes.

- The correlation coefficient is designed to summarize the linear relation between a pair of continuous variables that have approximately normal distributions. You can use a correlation to summarize the relations between variables with different distributions, but this makes it more difficult to detect a relation. This is particularly a problem when one or both of your variables are dichotomous.

  Sometimes you expect that the underlying nature of a variable is continuous even though it is measured in a categorical fashion. For example, people will often perform a median split on a variable in order to use it in an ANOVA design. In this case you can actually correct any observed correlations with that variable for the influence of the dichotomy using the formula

  $$r\{\text{dichotomy corrected}\} = r\{\text{observed}\}\frac{\sqrt{PQ}}{h}, \tag{7.1}$$

  where P and Q are the proportions of observations falling into the two categories of the dichotomous variable (so $Q = 1-P$), and h is the height of the normal distribution at the point where the probability to the left of the Z is equal to either P or Q (the heights will be the same at both points). Values of h can be obtained from Appendix C of Cohen, et al. (2003), or can be directly computed using the equation

  $$h = \frac{\exp\left(-\frac{Z^2}{2}\right)}{\sqrt{2\pi}}, \tag{7.2}$$

  where "exp" refers to the exponential function and Z is the value of the standard normal distribution where the probability to the left is equal to P.

- Correlation coefficients can be reduced if you have random error in the measurement of either variable. The *reliability* of a measure is defined as the proportion of the variability in the observed scores that can be attributed to systematic elements of the measure. Reliability ranges from 0 to 1, where higher values indicate more reliable measures. The maximum correlation that you can obtain with a measure is equal to the square root of the reliability. You can correct the observed correlation to determine what the relation would have been if the study had used a perfectly reliable measurement using the formula

$$r\{\text{reliability corrected}\} = \frac{r\{\text{observed}\}}{\sqrt{r_{xx} r_{yy}}}, \tag{7.3}$$

  where r_{xx} and r_{yy} are the reliabilities of your two variables.

- Correlations can also be reduced if you have a *restriction of range* in either of your variables. You can get a better estimate of the relation between a pair of variables when you examine them across a broader range. If you only have data from a limited range of one of your variables it can reduce your observed correlation. You can correct the observed correlation to determine what the relationship would have been if the study did not suffer from a restriction of range using the formula

$$r\{\text{range corrected}\} = \frac{r\{\text{observed}\}\left(\frac{s\{\text{full}\}}{s\{\text{observed}\}}\right)}{\sqrt{1 + r^2\{\text{observed}\}\left[\left(\frac{s^2\{\text{full}\}}{s^2\{\text{observed}\}}\right) - 1\right]}}, \tag{7.4}$$

  where s_{full} is the expected standard deviation of the variable when it does not suffer from restriction of range, and s_{observed} is the observed standard deviation of the variable suffering from restriction of range.

- You should always be very cautious when interpreting a correlation where one of the variables is a composite of other variables. Some examples of composites would be difference scores or ratios. If any of the individual components making up the composite are related to the other measure in the correlation, you will observe a correlation between the entire composite and the measure.

- See Hunter and Schmidt (1990) for a more complete list of other biases that may influence effect sizes as well as suggestions for how to correct these biases.

7.5 Multiple Dependent Variables Within Studies

- A study may sometimes use several different dependent variables to measure a single theoretical construct. You can deal with this situation in three ways.

  1. Calculate effect sizes for each response measure and enter them all in the same model. This is the easiest route, but it violates the independence assumption made by our analyses.
  2. Calculate effect sizes for each response measure and perform a separate analysis on each measure. This is really only feasible if the each response measure was used in a number of different studies.
  3. Mathematically combine the two effect sizes into one. This is the most preferred method.

- To combine effect sizes, meta-analysts often take a mean or median of the effect sizes computed separately on each response measure. This procedure is actually conservative if the response measures are correlated. It produces an estimate that is lower than one that would be produced from a test on a composite index of the response measures.

- Rosenthal and Rubin (1986) present a method for computing more accurate combinations of effect sizes.

- To combine several correlations you can use the formula

$$\text{combined } z_r = \frac{\sum z_{ri}}{\sqrt{\rho m^2 + (1-\rho)m}}, \tag{7.5}$$

where z_{ri} is the z transform of the correlation for the ith measure, ρ is the typical intercorrelation between the response measures, and m is the number of response measures you are combining.

- To combine several g statistics you can use the formula

$$\text{combined } g = \frac{\sum g_i}{\sqrt{\rho m^2 + (1-\rho)m}}, \tag{7.6}$$

where g_i is the effect size for the ith measure, ρ is the typical intercorrelation between the response measures, and m is the number of response measures you are combining.

One problem with using Rosenthal and Rubin's (1986) equations is that they require the typical intercorrelation between the response measures. You can seldom find this in every study in which you wish to combine effect sizes, but you can probably find it in some studies. You can use the correlations that you are able to get to estimate the correlations in the studies that do not provide them.

Chronbach's alpha (which is often reported) can be used to determine the average interitem correlation using the formula

$$\bar{r}_{ij} = \frac{\alpha}{n + (1-n)\alpha}, \tag{7.7}$$

where α is Chronbach's alpha and n is the number of response measures. $\bar{r}_{ij}$ can then be used for ρ in equations 7.6 and 7.5.

Chapter 8

Describing Effect Size Distributions

8.1 Introduction

- The methods for analyzing effect sizes are the same no matter what exact definition (i.e., mean difference, correlation, etc.) you decide to use. All formulas in this chapter and the next will therefore be written in terms of a generic effect size T. If you are working with mean differences, T would be equal to g^*. If you are working with correlations, T would be equal to Z_r.
- The first step to meta-analyzing a sample of studies is to describe the general distribution of effect sizes. A good way to describe a distribution is to report
    1. the center of the distribution
    2. the general shape of the distribution
    3. significant deviations from the general shape
- You should closely examine any outlying effect sizes to ensure that they are truly part of the population you wish to analyze. There are three common sources of outliers.
    1. The study methodology contains elements that alter the constructs being tested, such as when an irrelevant variable is confounded with the critical manipulation. These studies should be marked for exclusion from analysis.
    2. The outlier is the result of a statistical error by the original authors. If you suspect a statistical error you should mark the study for exclusion from analysis.
    3. The study tests the effect under unusual conditions or in nonstandard populations. You should endeavor to include these studies in analysis, since they are truly part of your population and provide unique information. You may wish to develop a code to examine the unusual condition as a moderator.

    If you cannot decide whether or not a given observation is an outlier you an run your analysis with and without the observation. If there are no differences you should keep the observation and report that dropping it would not influence the results. If there are differences you can choose to exclude the observation or report both analyses.
- If you have an important moderator that has a strong influence on your effect sizes, you might consider performing separate descriptive analyses on each subpopulation.

8.2 Nonstatistical Ways of Describing the Sample

- You can learn a lot about the distribution by examining a *histogram* of your effect sizes. A histogram plots effect size values on the x-axis and frequencies on the y-axis. Some of the most informative features of a histogram are

1. The number of modes in the distribution. Different modes could indicate the presence of significantly different subpopulations.
2. The overall shape of the distribution. You should consider whether your effect sizes appear to have a symmetric or skewed distribution.
3. The existence of outlying effect sizes. Any observations that appear to violate the general form of the distribution should be examined to determine whether they should be removed as outliers.

- It is typical to report the modal characteristics of your sample. You should therefore calculate the most common value of each of your moderators and then report them as the "typical" characteristics of studies in your sample.

8.3 Statistical Ways of Describing the Sample

- The goal of a statistical description of your sample is to provide information about the central tendency and variability of your collection of effect sizes. The models for meta-analysis have many similarities to the models used in primary research. However, they take several unique features of meta-analytic data into account, and so are somewhat different. It is inappropriate to use procedures designed to analyze primary research to draw conclusions about meta-analytic data.

- You can do this nonparametrically by presenting the median effect size along with the effect sizes corresponding to key percentiles of the distribution. One common method would be to present the "five-number summary," which would include the effect sizes

    - The mininum
    - The 25th percentile
    - The median
    - The 75th percentile
    - The maximum

  You might also choose to provide this information graphically in a boxplot.

- You can provide the weighted mean effect size and sample heterogeneity. Weights are used so that studies with larger sample sizes provide a greater contribution to the mean than those with smaller sample sizes. The exact weight that is used for each study depends on whether you decide to use a fixed- or random-effects model (see section 8.4 below), but will be directly related to sample size. You may also want to report the mean weighted effect size excluding studies in which you assumed that the effect size was zero because the study only reported that the test was nonsignificant, since these are the least accurate effect size estimates.

  Heterogeneity is a statistic that measures variability in a way that is similar to the standard deviation, but which takes the size of the study into account.

- You can provide the unweighted mean and standard deviation of your effect sizes. In this case you just calculate the mean and standard deviation of the observed effect sizes without worrying about the study sample sizes.

8.4 Choosing Between Fixed-Effects and Random-Effects Models

- Before you can analyze your effect sizes you have to decide whether you will base these analyses on a fixed-effects model or a random-effects model. This affects both the way you calculate your statistics as well as the types of inferences that you can draw from your results.

- The difference between the two models is how the study factor is defined in your model. Fixed-effects models define study as a fixed effect, whereas random-effects models define study as a random effect.

- Conceptually, a fixed-effects model assumes that all of your studies are estimating a single, common effect size. Each study acts as an additional estimate of mean effect size of the population. Study-to-study variability is assumed to arise solely from inaccuracies in measurement. If each study had very large sample sizes, the assumption is that their effect sizes would all converge because the population value for all of the studies is the same.

  A random-effects model, however, suggests that different studies are actually estimating slightly different effects. The population effect sizes are believed to form a distribution around the mean effect size. While some study-to-study variability represents measurement error, some represents true differences among the studies. If each study had very large sample sizes, the assumption is that their effect sizes would still be different because each study has a different population value. Random-effects models specifically assume that the collection of studies in your sample is a random selection from the underlying population.

- Fixed-effects models provide more powerful tests and smaller confidence intervals around the estimated mean effect size than random-effects models.

- A fixed-effects model allows you to generalize your results only to studies identical to those in your sample. In practice, "identical to" is typically interpreted as "quite similar to." Your inferences would be invalid if applied to situations with characteristics that weren't in your original sample, as well as those that combine characteristics present in your sample in unique ways.

  A random-effects model assumes that your the studies you observed are a random sample from a broader population of studies, allowing you to generalize to the population from which the sample was drawn. This allows you to generalize your results beyond the levels in your sample, as long as they can be thought of as belonging to the same population. It would still be invalid to draw inferences about situations that are notably different from those found in the studies contained in your sample.

- In most situations it random-effects models are more appropriate than fixed-effects models. Despite this, fixed-effects models have historically been more popular than random-effects models because the mathematics for fixed-effects models are simpler than those for random-effects models. Advances in statistical software, however, have made it considerably easier to use random-effects models. Most journals now expect meta-analysts to work with random-effects models unless they offer a strong justification for a fixed-effects model. For more information on the differences between fixed- and random-effects models see Hedges and Vevea (1998).

8.5 Using a Fixed-Effects Model to Describe a Sample

- Meta-analysts most often use the weighted average effect size when reporting the central tendency of their sample of studies. This may be calculated using the formula

  $$\bar{T} = \frac{\sum w_i T_i}{\sum w_i}, \tag{8.1}$$

  where

  $$w_i = \frac{1}{\text{variance of } T_i}. \tag{8.2}$$

  Some meta-analysts suggest setting w_i equal to the sample size instead of the inverse of the variance, though the latter is used much more often.

- The variance of the weighted average effect size may be calculated using the formula

  $$s^2_{\bar{T}} = \frac{1}{\sum w_i}. \tag{8.3}$$

- Using these statistics we can construct a level C confidence interval for θ (the population effect size):

  $$\bar{T} \pm z^*(s_{\bar{T}}), \tag{8.4}$$

where z^* is the critical value from the normal distribution such that the area between $-z^*$ and z^* is equal to C.

- We can also test whether θ (the population effect size) = 0 using the statistic

$$z = \frac{\bar{T}}{s_{\bar{T}}}, \tag{8.5}$$

where z follows the standard normal distribution.

- One important question is whether or not there is a common population effect size for the observed sample. To test the null hypothesis that all the studies come from the same population you can calculate the *heterogeneity statistic*

$$Q_T = \sum \left[w_i(T_i - \bar{T})^2\right] = \sum w_i(T_i)^2 - \frac{\left(\sum w_i T_i\right)^2}{\sum w_i}, \tag{8.6}$$

which follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of effect sizes in your sample. Large values of Q_T indicate that your observed studies likely come from multiple populations. Q_T can be interpreted as a comparison of between-study to within-study variance (Hedges & Vevea, 1998).

The statistic Q_T has been referred to as both "heterogeneity" and as "homogeneity" throughout the meta-analytic literature. We will call it "heterogeneity" in these notes because higher values of Q_T indicate that there is more diversity among your studies.

- Another measure of the dispersion of the distribution is the proportion of non-homogeneous effect sizes. To determine this you

  1. Calculate the heterogeneity Q_T of your distribution using equation 8.6.
  2. If Q_T is significantly different from zero, then you do not have a homogeneous distribution. You should remove the effect size farthest from the mean, then recalculate the heterogeneity.
  3. Continue dropping extreme studies until you have a homogeneous distribution.
  4. Count how many studies you had to drop to achieve homogeneity, and report the corresponding proportion.

It typically takes the removal of around 20% of the studies to get homogeneity. It may be informative to report the central tendency of the irrelevant part of the distribution.

8.6 Using a Random-Effects Model to Describe a Sample

- Just as with a fixed-effects model, researchers using a random-effects model will typically provide a weighted mean effect size when reporting the central tendency of a meta-analytic sample. The same formula (equation 8.1) is used to calculate the weighted mean effect size, but the weights used in a random effects model are different from those used in a fixed effects model.

- As mentioned in section 8.4, random-effects models assume that some of the study-to-study variability is due to measurement error, but some is due to actual differences among the studies. We treat these two sources differently when calculating the weighted mean effect size using a random-effects model.

- The first thing we must do to determine the weights for the mean effect size is calculate the variance attributable to study differences τ^2. This can be done by taking the following steps.

  1. Calculate the total heterogeneity of the sample using the equation

$$Q_T = \sum \left[w_i(T_i - \bar{T})^2\right] = \sum w_i(T_i)^2 - \frac{\left(\sum w_i T_i\right)^2}{\sum w_i}, \tag{8.7}$$

where

$$w_i = \frac{1}{\text{variance of } T_i}. \tag{8.8}$$

This is the same definition for heterogeneity that was used for fixed-effects models. Note that the weights w_i that we use in equation 8.7 will *not* be the same weights that we will use when calculating the weighted mean effect size.

2. Determine heterogeneity due to measurement error. If we assume that the measurement error follows a normal distribution, then we can calculate the expected value of Q_{error} using the equation

$$Q_{\text{error}} = \text{number of studies in the sample} - 1. \tag{8.9}$$

3. Determine the heterogeneity due to study differences. By definition, the total heterogeneity of a sample is the sum of the heterogeneity attributable to measurement error and the heterogeneity attributable to study differences, represented in the equation

$$Q_T = Q_{\text{study}} + Q_{\text{error}}. \tag{8.10}$$

We can rearrange the terms to obtain the formula for calculating the heterogeneity attributable to study differences:

$$Q_{\text{study}} = Q_T - Q_{\text{error}}. \tag{8.11}$$

There are times when this calculation can lead to estimates of Q_{study} that are less than zero. In these circumstances it is conventional to set $Q_{\text{study}} = 0$ because it does not make sense to have a negative number for heterogeneity.

4. Calculate the value of τ^2. τ^2 is directly related to Q_{study}, but needs to be on the same scale as our effect size variances so that they can be combined when calculating the weights. We can calculate τ^2 from Q_{study} using the equation

$$\tau^2 = \frac{Q_{\text{study}}}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}, \tag{8.12}$$

where w_i is calculated for each effect size using formula 8.8 above.

- Once you have τ^2, you can calculate the random-effects weight for each study using the equation

$$w_i^* = \frac{1}{s_{T_i}^2 + \tau^2}, \tag{8.13}$$

where $s_{T_i}^2$ is the effect size variance for that study.

- The weighted mean effect size is calculated as

$$\bar{T}^* = \frac{\sum w_i^* T_i}{\sum w_i^*}. \tag{8.14}$$

- The main difference between the fixed-effects weighted mean $\bar{T}$ and the random-effects weighted mean $\bar{T}^*$ is that studies with smaller sample sizes will have a larger influence on $\bar{T}^*$ than they will on $\bar{T}$. $\bar{T}^*$ will also have a larger standard error than $\bar{T}$.

- The variance of $\bar{T}^*$ can be calculated using the equation

$$s_{\bar{T}^*}^2 = \frac{1}{\sum w_i^*}. \tag{8.15}$$

- Using these statistics we can construct a level C confidence interval for θ (the population effect size):

$$\bar{T}^* \pm z^*(s^*_{\bar{T}}), \tag{8.16}$$

where z^* is the critical value from the normal distribution such that the area between $-z^*$ and z^* is equal to C.

- We can also test whether θ (the population effect size) = 0 using the statistic

$$z = \frac{\bar{T}^*}{s^*_{\bar{T}}}, \tag{8.17}$$

where z follows the standard normal distribution.

- In addition to reporting the weighted mean effect size as a measure of central tendency, you can also report Q_T and τ^2 as measures of effect size variability. Q_T follows a chi-square distribution with degrees of freedom equal to (number of studies in the sample - 1). A significant Q_T statistic would indicate that τ^2 is significantly greater than 0, indicating that there is more variability between studies than we would expect due to chance alone.

8.7 Interpreting Effect Sizes

- You should always put effort into interpreting the observed effect sizes for your audience. This will help give your readers an intuitive understanding of your results.

- When interpreting standardized mean difference effect sizes (Δ, d, g, or g^*), you can treat the different estimates as equivalent, even though there are minor differences between them. Most of the methods we present below were originally developed for d, but they can also be used for Δ, g, or g^*.

- If other meta-analyses have been performed in related topic areas, you can report the mean size of those effects to provide context for the interpretation of your effect.

- If no other meta-analyses have been performed on related topics you can compare the observed effect size to Cohen's (1992) guidelines:

Size of effect	Δ, d, g, or g^*	r
small	.2	.1
medium	.5	.3
large	.8	.5

Cohen established the medium effect size to be one that was large enough so that people would naturally recognize it in everyday life, the small effect size to be one that was noticeably smaller but not trivial, and the large effect size to be the same distance above the medium effect size as small was below it.

- You can provide a measure of how certain you are that your effect is not caused by publication bias by reporting the number of unreported studies with null findings there would have to be so that your mean effect size would not be significantly different from zero. This number may be calculated (using the Stouffer method of combining probabilities) from the equation

$$X = \frac{\left(\sum z_i\right)^2}{2.706} - N_L, \tag{8.18}$$

where z_i is the z score associated with the 1-tailed p value for study i, and N_L is the total number of located studies. In his description of this method, Rosenthal (1991) provides a somewhat arbitrary comparison point, claiming that if $X > 5N_L + 10$ then it is implausible that an observed significant effect size is truly nonsignificant. This method assumes that the mean effect size in unreported studies is zero, which may not be true if the publication bias favors outcomes in one tail of the distribution.

- To help give intuitive meaning to an effect size you can provide the Binomial Effect Size Display (BESD), presented in Rosenthal and Rubin (1982). This index presents the proportion of cases (or people) who succeed in the experimental group and the proportion that succeed in the control group. The definition of "success" is based on the way your effect size is defined. This index makes the most sense when researchers use status on a dichotomous predictor variable (such as experimental vs. control group) to predict a dichotomous outcome (such as succeeding versus failing). The easiest way to calculate the BESD is to

  1. Transform your effect size statistic into r
  2. Calculate the success rate of your experimental group

  $$\text{success}_E = .5 + \frac{r}{2} \tag{8.19}$$

  3. Calculate the success rate of your control group

  $$\text{success}_C = .5 - \frac{r}{2} \tag{8.20}$$

  The BESD counters the tendency for people to trivialize small effects. For example, a researcher might conclude that an correlation of .2 is small because it accounts for only 4% of the variance. The BESD allows you to realize that, nonetheless, people's success rates would be 20% higher in the experimental group than in the control group.

  When the response variable is continuous you must dichotomize it at the median to interpret this index. You would then conclude that a person would have a probability of $.5 + \frac{r}{2}$ of being above average in the experimental group and a probability of $.5 - \frac{r}{2}$ of being above average in the control group.

- You can also use the Common Language effect size statistic (CL), presented in McGraw and Wong (1992), to help you interpret your effect size. This index is the probability that a score randomly drawn from one distribution will be larger than a score randomly drawn from another distribution. Dunlap (1994) provides a method to convert the effect size d to CL, allowing you to report the probability that a person drawn randomly from the experimental group provides a greater response than a person drawn randomly from the control group.

- Cohen (1977) provides three additional measures that can help you interpret an effect size.

  - U_1: the percent of the total area covered by the distributions in the experimental and control groups that is non-overlapping. To calculate U_1 you
    1. Look up $\frac{d}{2}$ in a normal distribution table and record the area to the right as A and the area to the left as B.
    2. Calculate the nonoverlap of the experimental group $C = B - A$.
    3. Calculate the total nonoverlap $U_1 = \frac{2C}{2B}$.
  - U_2: the percent of the experimental population that exceeds the same percentage in the control population. This is the proportion of the group with the larger mean effect size that is to the right of the point where the two distributions cross. To calculate U_2 you look up $\frac{d}{2}$ in a normal distribution table. U_2 will be the percentage to the left.
  - U_3: the percent of those in the experimental group that exceed the average person in the control group. To calculate U_3 you look up d in a normal distribution table. U_1 will be the percentage to the left.

  Of these measures, U_3 is used most often because of the ease of its interpretation.

8.8 Vote-Counting Procedures

- Vote-counting solely uses information about the direction of findings to generate conclusions about the literature. These procedures can be useful as a secondary technique of looking at studies' findings. Also, when the information required to calculate effect sizes is missing from many studies in your sample, you may have to revert to weaker vote-counting methods to aggregate effects across studies.
- The accepted method calculates the exact p value of the obtained distribution of outcomes (or one more extreme), given that the true population effect size = 0. The calculations rest on the assumption that any single study has a .5 probability of a positive result and a .5 probability of a negative result under the null hypothesis. This is referred to as the "sign test," and based on the binomial distribution.
- To perform the sign test you

    1. Determine the probability that you get a positive result. In the situation above this would be .5.
    2. Count the total number of studies and assign this value to n.
    3. Count the number of studies that give positive results and assign this value to m.
    4. Look up the probability in a cumulative binomial probability distribution. Most statistical software packages (including SPSS and SAS) have functions that will also provide you with the appropriate probability.

  You can interpret the resulting p value in the standard way, testing whether $\theta = 0$.

Chapter 9

Moderator Analyses

9.1 Overview

- Moderator analysis allows you to determine whether study characteristics are significantly related to effect sizes. If a moderator is categorical, you can test whether the mean effect sizes for the different groups are significantly different from each other. If a moderator is continuous, you can test whether there is a significant linear relation between a the value of that moderator and the study effect size. You can also use a parallel to multiple regression analysis to determine the joint and unique abilities of a collection of moderators to explain variability in the effect sizes.

- Meta-analysts will sometimes base the decision to perform moderator analyses on effect size heterogeneity, which is calculated using equation 8.6. By definition, a significant heterogeneity test indicates that there is more variability among the effect sizes than you would expect due to random chance alone. One possible reason for this heterogeneity is that the moderator variables are influencing the strength of the effect between studies, suggesting the need for moderator analyses.

  While a significant heterogeneity statistic does suggest that moderator analysis may prove fruitful, the test of heterogeneity is not very powerful when there are a small number of studies in the sample. In this case it may be reasonable to conduct moderator analyses even when the heterogeneity statistic is not significant.

- Before starting your moderator analyses, you must determine whether you will treat the study factor as a fixed or random effect in your model. The way you treat the study factor in your moderator analyses should be the same way you treated it when estimating the weighted mean effect size. The moderator variable itself is usually treated as a fixed effect.

  Models that test a moderator while defining study as a random effect are typically referred to as "mixed models" because they include both fixed and random factors. In this chapter we will provide specific details about how to test moderators using fixed-effects models, but we will not provide details about how to test moderators using mixed models. The tests are conceptually the same in a mixed model, but it requires matrix algebra or iterative methods to estimate the statistics, which goes beyond the scope of these notes. Details on testing moderators in mixed models can be found in Overton (1998). Lipsey and Wilson (2000) provide SPSS and SAS macros to perform moderator analyses using mixed models.

- Next you determine the ability of each moderator to explain variability in the effect sizes. This can be done for categorical moderators by determining the between-group and within-group heterogeneity. It can be done for continuous moderators by estimating the slope and standard error of the linear relation between the moderator and effect sizes. These initial analyses typically examine a single moderator at a time.

- After determining which moderators can explain variability in the effect sizes, you will typically examine the relations among the moderators. It is important to know if any of the observed bivariate relations might actually be caused by confounding between the moderators.

- Multiple regression can then be used to explore more complicated models predicting the effect sizes. This can be used to determine the unique predictive ability of moderators that are related to each other, to examine nonlinear relations between moderators and the effect sizes, and to explore the effects of interactions between moderators.

9.2 Testing a Categorical Moderator

- In primary research we often use ANOVA to assess the ability of a categorical predictor variable to explain a numeric response variable. The heterogeneity statistic Q_T presented in equation 8.6 is a measure of the total variability within a set of effect sizes. Similar to the way we partition variance when performing an ANOVA, we partition the variability represented by Q_T when performing a meta-analysis.
  - When using a fixed-effects model, Q_T will be partitioned into Q_B (between-groups heterogeneity), the part explained by the moderator, and Q_W (within-groups heterogeneity), the part that is not explained by the moderator.
  - When using a mixed model, Q_T will be partitioned into Q_B, the part explained by the moderator, Q_{study}, the part explained by the random effect of study, and Q_W, the part not explained by either.

- We use the between-groups heterogeneity Q_B to measure how much variability can be explained by a moderator. To calculate Q_B in a fixed-effects model, you

  1. Calculate the weighted mean and variance of the effect sizes for each level of your moderator using equations 8.1 and 8.3.
  2. Calculate Q_B using the equation

  $$Q_B = \sum w_i(\bar{T}_i - \bar{T})^2, \tag{9.1}$$

  where $\bar{T}_i$ is the mean of group i, the weight is calculated as

  $$w_i = \frac{1}{s^2_{\bar{T}_i}}, \tag{9.2}$$

  where $s^2_{\bar{T}_i}$ is the variance of the mean effect size for level i and the summation is over all of the levels of the moderator.

  The statistic Q_B follows a chi-square distribution with $p-1$ degrees of freedom, where p is the number of levels in your moderator. Large values of Q_B indicate that your moderator can predict a significant amount of the variability contained in your effect sizes.

- We use the within-groups heterogeneity Q_W to measure how much variability the moderator fails to explain. To calculate Q_W in a fixed-effects model you

  1. Calculate the variability within each individual level of your moderator. The variability Q_{W_j} for level i is simply the heterogeneity (see equation 8.6) of the studies within level i.
  2. Calculate Q_W using the equation

  $$Q_W = \sum {Q_W}_i, \tag{9.3}$$

  where the summation is over the different levels of the moderator.

  The statistic Q_W follows a chi-square distribution with $k-p$ degrees of freedom, where k is the total number of effect sizes in your sample. Large values of Q_W indicate that there is a significant amount of variability in your effect sizes that is *not* accounted for by your moderator.

- Q_B and Q_W partition the total heterogeneity in a fixed-effects model. That is,

$$Q_T = Q_B + Q_W. \tag{9.4}$$

- When your categorical model contains more than two groups you will probably want to compute contrasts to compare the group means. Rosenthal and Rubin (1982) show that you can test a contrast by taking the following steps.

  - Define your contrast as a linear combination of the group mean effect sizes, using the form

  $$L = \sum_{i=1}^{p} c_i \bar{T}_i, \tag{9.5}$$

  where p is the number of levels in your moderator, c_i is the contrast weight for level i, and $\bar{T}_i$ is the mean effect size for level i. The sum of your contrast weights must always be equal to 0.
  - To test whether L is significantly different from zero you would calculate the test statistic

  $$Z = \frac{L}{\sqrt{\sum\limits_{i=1}^{p} \frac{c_i^2}{w_i}}}. \tag{9.6}$$

  p-values for Z can be drawn from the standard normal distribution.

9.3 Testing a Continuous Moderator

- You can test whether there is a linear relation between a continuous moderator and your effect sizes using procedures analogous to simple linear regression for primary data. Detailed information about meta-analytic regression procedures is presented in Hedges (1994).

- To test a continuous moderator you

  1. Transport your meta-analytic database into a standard computer package (SAS, SPSS).
  2. Create a variable equal to the reciprocal of the variance (if you hadn't already created it at some prior stage in your analysis).
  3. Perform a weighted regression using the reciprocal of the variance as the case weight.
  4. Draw your regression coefficients directly from the output.
  5. Calculate the standard deviation of the slope (which is *not* equal to that provided on the output) using the equation

  $$s_{b1} = \frac{u_{b1}}{\sqrt{\text{MSE}}}, \tag{9.7}$$

  where u_{b1} is the (incorrect) standard error of the slope provided by the computer program and MSE is the mean square error of the model.
  6. Calculate the test statistic

  $$Z = \frac{b_1}{s_{b1}}, \tag{9.8}$$

  which follows the standard normal distribution. Large values of Z indicate that there is a significant linear relation between effect size and your moderator.

9.4 Examining Relations Among Moderators

- Very rarely will you find that the moderators in your study are independent: There will almost always be covariation in study features and manipulations. These covariations can provide valuable insights into the character of your literature.
- At its heart, a meta-analysis is really just an observational study. Like any non-experimental design, to establish that there is a causal relation between two variables (such as a moderator and the effect size) you need to not only show that a relation exists between the two but also that the relation is not the caused by the action of a third variable.

  Practically, if two moderators are highly correlated and the first causes changes in the effect size, a moderator test for the second will likely also be significant even though it does not truly influence the strength of the effect. Other types of relations between moderators can cause the test of an important moderator to be nonsignificant. It is difficult to draw any strong conclusions about correlated moderators, so when possible it is best to define your moderators in such a way so that they are not correlated.
- To test the relations among your moderators you will want to create a data set that has one case for each study and which has variables representing the values on the different moderators. Studies that manipulate the values of a moderator should be excluded from the analyses of relations with that moderator.
- To examine the relation between two categorical moderators you can create a *two-way table.* In a two-way table, the values of one moderator are placed on the horizontal axis, the values of the second moderator are placed on the vertical axis, and the inside of the table reports the number of studies you have with that particular combination of variables. You can perform a chi-square test to see if there is a significant relation between your two moderators.
- The easiest way to examine the relation between two continuous moderators is to calculate the Pearson correlation. If you want to test for a more complicated relation (such as quadratic) you can use regression analysis and test the values of the coefficients.
- To examine the relation between a categorical moderator and a continuous moderator you can calculate the mean value of the continuous moderator at each level of the categorical moderator. You can test the strength of the relation using ANOVA.
- You might consider weighting each study in these analyses by the sample size, or might want to assign equal weight to each study regardless of the sample size. Either choice is defensible, but your decision will influence the meaning of your results. Weighted analyses provide information about the covariation of conditions by subject, while unweighted analyses provide information about the covariation of conditions by study.
- Since most meta-analyses code a large number of characteristics, it may not be feasible to test for covariation between every pair of moderators. It is therefore common practice to only examine the relations among moderators that are significantly related to the effect size. You may also choose to test any specific relations that you feel would be particularly interesting to examine.
- When presenting the relations between moderators in your analysis, you may choose to use a single statistic (such as correlations) to make it easier to compare the strength of the relations. In this case, you would convert all measures of association that you calculate into the chosen form.

9.5 Multiple Regression in Meta-Analysis

- In addition to testing whether effect sizes are related to the values of a single moderator, you can use multiple regression to perform more complicated analyses. Some examples are
  - Testing models with more than one moderator.

- ○ Testing for interactions between moderators.
- ○ Testing higher-order polynomial models.

- It is also becoming common practice to follow up a set of moderator analyses with a multiple regression model containing all of the significant predictors. The multiple regression model provides a control for the total number of tests, reducing the likelihood of a Type I error. It also helps you to detect whether collinearity might provide an alternative explanation for some of your significant results.

- The procedure for multiple regression closely parallels that for testing a continuous model:

  1. Transport your meta-analytic database into a standard computer package.
  2. Create a variable equal to the reciprocal of the variance.
  3. Create dummy variables for any categorical moderators. For more information about working with dummy variables see Hardy (1993).
  4. Perform a weighted regression using the reciprocal of the variance as the case weight.
  5. Draw your regression coefficients directly from the output.
  6. Calculate the standard deviations of your coefficients using the equation

  $$s_{bj} = \frac{u_{bj}}{\sqrt{\text{MSE}}}, \tag{9.9}$$

     where u_{bj} is the standard error of b_j provided by the computer program and MSE is the mean square error of the model.

- You can test and interpret the parameter estimates of your model just as you typically do in multiple regression. Recall that tests on the individual parameters examine the unique contributions of each predictor. You should therefore be careful to consider the possible effect of multicollinearity on your parameter estimates. You can use the procedures described in section 9.4 to see if multicollinearity might be a problem.

- You can also perform an overall test of your model. You can divide the total variability in your effect sizes (Q_T) into the part that can be accounted for by your model (Q_B) and the part that cannot (Q_E).

  - ○ Q_B is estimated by the sum of squares regression of your model, which can be taken directly from your computer output. It follows a chi-square distribution with p degrees of freedom, where p is the number of predictor variables (not including the intercept) included in your model. Large values of Q_B indicate that your model is able to account for a significant amount of the variance in your effect sizes.
  - ○ Q_W is estimated by the sum of squares error of your model, which also can be taken directly from your computer output. It follows a chi-square distribution with $k-p-1$ degrees of freedom, where k is the number of effect sizes in your analysis. Large values of Q_E indicate that your model does not completely account for the variance in your effect sizes.

Chapter 10

Writing Meta-Analytic Reports

10.1 General Comments

- One of the reasons that researchers developed meta-analysis is to provide a way of applying the scientific methods used in primary research to the process of reviewing. The steps to performing a meta-analysis therefore have some fairly direct parallels to the steps of primary research.
- The easiest way to write up a meta-analysis is to take advantage of this parallel structure by using the same sections found in primary research. When writing a quantitative literature review you should therefore include sections for the Introduction, Methods, Results, and Discussion.

  You need to present this same information when reporting a meta-analytic summary, though not always using the same format. If your summary includes moderator analyses, you should present it as a separate study in your paper, using the guidelines for reporting a quantitative review described above. However, if you are only presenting descriptive analyses, your meta-analysis will likely be simple enough that you can incorporate it directly into your introduction or discussion. In this case you should describe the purpose and method of your meta-analysis in one paragraph, with the results and discussion in a second.
- Overall, you should try to make your report as complete and clear as possible. In each section you should state every decision that you made that affected the analysis, and you should describe it in as plain terms as is possible.

10.2 Introduction

- Your introduction should concretely define the topic of your analysis and place that topic into a broader psychological context.
- To describe your topic area you should present
  - A description of the literature that you want to analyze in general terms, just as you would if you were writing a primary research article on the topic.
  - An explanation of why a meta-analysis is needed on your topic.
  - A discussion of theoretical debates in the literature.
  - Explanations of any unusual terminology or jargon that you will be using in the paper.
- To specify how you analyzed the literature you should present
  - A precise definition of the effect you are examining.
  - A theoretical description of the boundaries of the analysis. This should justify the inclusion/exclusion criteria that you will be using.

    - A description of any significant subgroups of studies found in the literature.
    - The theoretical background behind any statistical models you decided to test.

- You may also want to use your introduction to present the organization of the remainder of the paper, especially if you perform several sets of analyses.

10.3 Method

- In the method section you need to describe how you collected your studies and how you obtained quantitative codes from them.

- To describe how you collected your studies you should present

    - A thorough description of your search procedure including

        1. The name of each computer database you used, the search terms you used, and the years covered by the database.
        2. Review articles you searched for references.
        3. The names and volumes of journals you searched by hand.
        4. A description of any attempts you made to contact authors in search of unpublished work.

      You should describe your search procedures in such a way that other researchers could replicate your work.

    - The criteria you used to include and exclude studies from analysis. You might also decide to report examples to clarify the criteria.

- To describe how you coded moderator variables you should present

    - An explanation of your general coding method. You should report

        1. How many coders you used.
        2. How familiar the coders were with the literature being reviewed. This might include the degrees possessed by the coders and the amount of experience they have in the field.
        3. Whether you coded one or more than one effect from each study. If you coded multiple effects, you should report how you decided how many effects to code from each study.
        4. How you resolved differences between coders.

    - Descriptions of each moderator you coded. For each moderator you should explain

        1. Why you decided to include the moderator in your analysis.
        2. What units (for continuous moderators) or categories you used (for categorical moderators) in coding.
        3. The rules you used to code the moderator.
        4. The coding agreement rate.

- To describe how you calculated your effect sizes you should present

    - Definitions of the the variables composing the effect size. For mean differences, this would be the definition of the two groups you are comparing. For correlations, this would be the two variables involved in the correlation.
    - A general description of how the variables were commonly operationalized in the literature.
    - What different methods you used to calculate the effect size. If there are multiple ways to calculate the effect size you should report how you decided which one to apply in a given case.

- You should also describe any unusual issues you were forced to deal with during searching, coding, or the calculation of your effect sizes.

10.4 Results

- In the results section you describe the distribution of your effect sizes, present any moderator analyses you decided to perform, report the observed relations among the moderators, and present any multiple regression models you analyzed.
- To describe the distribution of your effect sizes you should present
  - A histogram of the effect sizes.
  - A discussion of possible outliers.
  - "The typical study" – a report of the modal moderator values.
  - Descriptive statistics including
    1. The number of studies included in the analysis.
    2. Total number of research participants.
    3. Mean weighted effect size with confidence interval.
    4. Range of effect sizes.
    5. The overall heterogeneity Q_T and its corresponding p-value.
- For each categorical moderator you want to test you should present
  - Descriptive characteristics of each level of the moderator including
    1. The number of effects included in the level.
    2. The number of research participants included in the level.
    3. The mean weighted effect size.
    4. The within-group heterogeneity Q_{Wj} and its corresponding p-value.
  - The between-group heterogeneity Q_B and its corresponding p-value.
  - The total within-group heterogeneity Q_W and its corresponding p-value.
  - Any contrasts you choose to perform to help interpret a significant moderator.
- For each continuous variable you want to test you should present
  - The slope coefficient b_1.
  - The standard error of the slope s_{b1}.
  - A significance test of the slope.
- To describe the relations among the moderators you should present
  - Which moderators you considered. If you only examined a subset of your moderators, you should explain why you chose to examine the particular set that you did.
  - Information about each bivariate relation, including
    1. The type of test you used.
    2. The resulting test statistic and degrees of freedom.
    3. The p-value for the test.

    You may want to present the results from similar tests all in one table.
  - The implications of the results for your moderator analyses. You should explicitly discuss whether the relations among the moderators might offer alternative explanations for observed relations between moderators and the effect size.
- To describe the ability of a multiple regression model to explain your distribution of effect sizes you should present

    - A justification for the variables that were included in the model. If you decided to test all of the significant moderators in your multiple regression model, you can just use that as your justification.
    - The variability accounted for by your model Q_B and its corresponding p-value.
    - The variability not accounted for by your model Q_E and its corresponding p-value.
    - Tests of each parameter in the model.

10.5 Discussion

- To help your audience interpret the mean effect size you can present
    - References to other established effect sizes.
    - Rosenthal's (1991) file-drawer statistic.
    - Other statistics mentioned in section 8.7 designed to provide intuitive meaning to effect sizes.
- You should attempt to provide an explanation for any significant moderators revealed by your analyses. Ideally you will be able to use a single theoretical perspective to explain a collection of your significant moderators.
- You should describe the performance of any models you built in attempts to predict effect sizes.
- You should discuss the diversity of the studies in your sample.
- You should consider the implications of your findings for the major theoretical perspectives in the area of analysis.
- You should make theoretical inferences based on your results. What implications might they have for applied settings?
- You should mention any features of your analysis that might limit the generalizability of the results.
- You should conclude with specific recommendations for the direction of future research.
    - You should highlight specific conditions under which the effect has only rarely been investigated. This might include particular populations or particular levels of your moderator variables.
    - You should discuss important sources of multicollinearity that you found, and suggest future studies that can better separate the variables.

10.6 Miscellaneous

- You should have a single reference section that includes both studies used in writing the paper and those included in the meta-analysis. You should place an asterisk next to those studies included in the analysis.
- You should prepare an appendix including all of the codes and effect sizes obtained in the analysis. Many journals will not be interested in publishing this information, but you will likely receive requests for it from people who read your report.

Chapter 11

Critically Evaluating a Meta-Analysis

11.1 Overview

- Just as there are high and low quality examples of primary research, there are high and low quality meta-analyses. The diversity may be even greater within meta-analysis, since many reviewers are not familiar enough with the procedures of meta-analysis to differentiate between those that are good and those that are poor.

- It is especially important to critically examine meta-analyses conducted in the early 1980s. Those conducted at that time were not subject to as rigorous evaluation by reviewers as they are today, mostly because meta-analytic techniques were not widely understood.

- A good meta-analysis uses appropriate methods of data collection and analysis (possesses internal validity), properly represents the literature being analyzed (possesses external validity), and provides a distinct theoretical contribution to the literature. The following sections provide some specifics to consider when evaluating each of these dimensions.

11.2 Internal Validity of the Analysis

- The first thing to examine is the internal validity of the primary research studies themselves. Ultimately, a meta-analysis can never be more valid than the primary studies that it examines. If there are methodological problems with the studies then the validity of the meta-analysis should be equally called into question.

- The meta-analysis should contain enough studies to provide power for its test. The exact number will depend on what analyses are being performed. You only need a few studies to examine the overall effect size, but you would typically want at least 30 studies to examine moderator variables.

- If a meta-analysis performs moderator tests it should also report if there are any relations between the moderators. You should critically examine all results involving correlated moderators to see if there is a logical reason to doubt the interpretation of the results.

- Today, all meta-analyses will have at least two authors to ensure coding reliability. The reliability should be published, and should be reasonably high, preferably over .8.

- Standard meta-analytic procedures assume that all of the effect sizes are independent. If an analysis includes more than a single effect size per study, this assumption is violated. Sometimes the designs of the primary studies will require this violation, but the authors should take steps to minimize its impact on their results.

- Random-effects models are typically preferred to fixed-effects models because of the likely dependence due to study. The authors should specifically justify their decision if they choose to work with a fixed-effects model. Fixed-effects models are more acceptable if they include procedures (such as those discussed by Cooper, 1989) to reduce the effect of dependence on their findings.

- Assumed 0 effect sizes from reported null findings are the least precise effects that can be calculated. You should be cautious when drawing inferences from a meta-analysis that contains a substantial amount of these effects. If there are a large number of assumed 0 effect sizes, the authors should report their results both including and excluding these values from their analyses.

11.3 External Validity of the Analysis

- Possibly the most important factor affecting the external validity of a meta-analysis is the representativeness of the sample of studies. Ideally the sample of a meta-analysis should contain every study that has been conducted bearing on the topic of interest. Barring this, the sample of studies should be a fair representation of the literature as a whole.

  To assess the representativeness of a particular meta-analysis you should ask

  1. Do the theoretical boundaries proposed by the authors make sense? Does the studies in the analysis actually compose a literature unto themselves? Sometimes they can be too broad, such that they aggregate dissimilar studies. Other times they may be too narrow, such that the scope of the meta-analysis is smaller than the scope of the theories developed in the area.
  2. Did the authors conduct a truly exhaustive literature search? You should evaluate the keywords they used to search computerized indices, and what methods they used to locate studies other than through computerized indices.
  3. Did the authors look in secondary literatures? While the majority of the studies will likely come from a single literature, it is important to consider what other fields might have conducted research related to the topic.
  4. Did the authors include unpublished articles? If so, how rigorous was the search? If they did not, do they provide a justification for this decision?

- Having a very large literature is no excuse for failing to conduct an exhaustive search. If there are too many studies to reasonably include them all in the analysis, a random sample should be selected from the total population.

- The effects calculated for each study should represent the same theoretical construct. While the specifics may be dependent on the study methodology, they should all clearly be examples of the same concept.

- If the analysis included high-inference coding, the report should state the specifics of how this was performed and what steps they took to ensure validity and reliability. All high-inference moderators deserve to be looked at closely and carefully.

11.4 Theoretical Contribution

- A meta-analysis should not simply be a summary of a literature, but should provide a theoretical interpretation and integration. In general, the more a meta-analysis provides beyond its statistical calculations the more valuable its scientific contribution.

- Miller and Pollock (1994) divides meta-analyses into three categories based on their purpose and the type of information that they provide.

  - Type A analyses summarize the strength of an effect in a literature. Its main goal is to determine whether or not a postulated effect exists, and to measure its strength.

  - Type B analyses attempt to examine what variables moderate the strength of an effect. In some cases this involves determining the circumstances where a difference is absent or present, while in others it involves locating factors that enhance or diminish the effect of some treatment.
  - Type C analyses attempt to use meta-analysis to provide new evidence in relation to a theory. It moves beyond examining the moderators proposed by those conducting the primary studies and introduces a new potential moderator. Often times the newly proposed moderator cannot be reasonably tested in primary research, such as author gender or nationality.

  Type A analyses can be seen to make the smallest theoretical contribution, followed by Type B and then Type C. While this is only a gross division (a well-conducted Type B analysis is definitely more valuable than a poorly-conducted Type C analysis, for example), it serves to highlight the fact a good meta-analysis provides more than a statistical summary of the literature.

- A good meta-analysis does not simply report main effect and moderator tests. It also puts effort into interpreting these findings, and presents how they are consistent or inconsistent with the major theories in the literature.
- Meta-analyses can greatly aid a literature by providing a retrospective summary of what can be found in the existing literature. This should be followed by suggestions of what areas within the literature still need development. A good meta-analysis encourages rather than impedes future investigations.

References

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Boston: Houghton Mifflin Company.

Cooper, H. M., (1989). *Integrating Research: A Guide for Literature Reviews* (2nd ed.). Newbury Park, CA: Sage.

Cooper, H., & Hedges, L. (1994). *The Handbook of Research Synthesis.* New York: Russel Sage Foundation.

DeCoster, J. (2005). Meta-analysis. In Kempf-Leonard, K. (Ed.), *The Encyclopedia of Social Measurement.* San Diego, CA: Academic Press.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper and L. V. Hedges (eds.) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

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

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in Social Research.* Beverly Hills, CA: Sage Publications.

Hedges, L. V. (1994). Fixed effects models. In Cooper, H., & Hedges, L. V. (Eds.) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Beverly Hills, CA: Sage.

Hardy, M. A. (1993). *Regression with Dummy Variables.* Sage University series on Quantitative Applications in the Social Sciences, series no. 07-094. Newbury Park, CA: Sage Publications.

Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis in social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social Psychology* (pp. 496-528). London: Cambridge University Press.

Kenny, D. A. (1979). *Correlation and Causality.* New York: Wiley.

Lipsey, M. W., & Wilson, D. B. (2000). *Practical Meta-Analysis.* Thousand Oaks, CA: Sage.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin,* *111*, 361-365.

Miller, N., & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H. Cooper and L. V. Hedges (eds.), *The Handbook of Research Synthesis.* New York, NY: Russell Sage Foundation.

Montgomery, D. C. (1997). *Design and Analysis of Experiments.* New York: Wiley.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105-125.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear Statistical Models.* Chicago: Irwin.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3*, 354-379.

Rosenthal, R., & Rubin, D. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 99*, 400-406.

Rosenthal, R., & Rubin, D. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99*, 400-406.

Rosenthal, R. (1991). *Meta-Analytic Procedures for Social Research.* Newbury Park, CA: Sage.