Meta-Analysis Notes

Jamie DeCoster

Department of Social Psychology
Free University Amsterdam
Van der Boechorststraat 1
1081 BT Amsterdam
The Netherlands

phone: +31 (0)20 444-8935
email: j.decoster@psy.vu.nl

April 10, 2003

These were compiled by Jamie DeCoster, partially from a course in meta-analysis taught by Alice Eagly at Northwestern University. Handbook references refer to Cooper & Hedges (eds.), *The Handbook of Research Synthesis*.

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." Providing a report of primary research using statistical methodology and analysis is called "quantitative synthesis" or "meta-analysis." A report of primary research using traditional, literary methods is called a "narrative review."
- Meta-analyses are generally centered on the relationship between one explanatory and one response variable. This relationship, "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.
- Meta-analyses are most easily performed with the assistance of computer databases (Microsoft Access, Paradox) and statistical software (DSTAT, SAS).

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 stronger simply but adding more participants.
- Narrative reviews lack acceptable rules of inference for going from the findings of studies to overall generalizations about the research literature.
- 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 to 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 relationship 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

1. Define the theoretical relationship of interest.
2. Collect the population of studies that provide data on the relationship.
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 the 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 your explanatory and response variables.
- You need to decide what type of effect size you will use. If the explanatory variable is typically presented as a categorical variable, you should probably use d. If the explanatory variable is typically presented as a continuous variable, you should probably use r.
- If you decide to use the effect size d, you then need to precisely define what contrast you will use to calculate d. For a simple design, this will probably be (mean of experimental group - mean of control group). Defining the contrast also specifies the directionality of your effect size (i.e., the meaning of the sign). The directionality is automatically determined for the effect size r once you choose your constructs.

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 define 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 relationship. Without this information there is nothing you can do with a study meta-analytically.
- 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 recheck studies you've already completed.
- It is a good practice to keep a list of the studies that turned up in your initial search but that you later decided to exclude from your analysis. You should also record exactly what criterion they failed to meet, so that if you later decide to relax a particular criterion you know exactly what studies you will need to re-examine, saving you from having to perform an entirely new literature search.

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 *master 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.
  3. Gain access to each of these studies for examination. Some of the studies will be available at your library, while others will have to be obtained either through interlibrary loan or directly from the authors.
  4. Examine each of the studies on this list and 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.
- You want to make sure that your master 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.
- You do not need to copy every study in the master candidate list. Many of these you will reject with just a few minutes of reading. However, you will want to copy each article in your final meta-analytic sample.
- 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 master 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 criterion was used for exclusion (if the study was excluded from the meta-analysis)

- 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 AltaVista Translation website (http://babelfish.altavista.digital.com/cgi-bin/translate?) 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)
  - 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. PsycLit/PsycInfo (PsycLit is the CD-ROM version, and is less complete)
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. 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 very nice complement to the standard computerized search, and can now be performed fairly easily since both indices are available on CD-ROM.

- *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, APS membership guides

Chapter 4

Coding Studies

4.1 How to Code

- Once you have collected your sample of studies you need to code their characteristics 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. Once you have a stable coding scheme you 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.
- 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 effect size calculations.
- Sometimes the information you need will not be reported in a study. You should therefore have a separate code 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, Chronbach's alpha, and (for categorical variables) Cohen's kappa.

- 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

- The first set of characteristics you need to code are study identifiers, including
  - *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 codes.
- You should code all moderating variables you wish to examine. The next section provides a detailed discussion of the different types of moderators you might wish to consider.
- You should code 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.
- You need to record information about the overall design of the study such as
  - *Assignment of subjects.* You should record whether subjects were assigned to conditions randomly or in some other fashion.
  - *Experimental design.* This should be an exact specification of the study design, specifying which factors are crossed and nested. You should specify all aspects of the design, not just those relevant to your analysis. Example: Time X S(Gender X Study method)
  - *Manipulation codes.* If important moderators are sometimes manipulated between subjects and sometimes within subjects you should code this as a moderating variable.
- You also need to record information about how you calculated the effect size. If you are using r you should report
  - *Correlation definition.* The variables, as defined in the study, that are used in the calculation of the effect size.
  - *Calculation method.* A code indicating what basic procedure you performed to get the effect size (directly reported, obtained from regression equation, etc.).
  - *N.* The number of subjects that were measured to calculate the correlation.

  If you are using d you should report
  - *Contrast definition.* A general description of the cells from the design that you used to calculate the difference for the effect size, as well as the response variable if it might be at all questionable. Example: (primed - unprimed hostility ratings for males)
  - *Dependent measure.* A verbal description of the dependent measure used to calculate the effect size.
  - *Calculation method.* An indication of the basic procedure you used to obtain the effect size. A sample set of codes might be
    1. Means and standard deviation
    2. t or 1 df F statistic
    3. Correlation coefficient

4. Proportions
5. χ^2 statistic
6. p-value
7. Combination of previously calculated effect sizes
8. Assumed effect size of 0 from reported null effect
9. No effect calculated

- *Source of means.* How you obtained the means to calculate the effect size (directly reported, estimated from a graph, calculated from other data, etc.), if means were used. A sample set of codes might be
  1. Directly reported
  2. Average of reported means
  3. Obtained from graph
  4. Means not used to calculate effect
- *Source of standard deviation.* How you obtained s_p to calculate the effect, if it was used. A sample set of codes might be
  1. Directly reported s's
  2. Directly reported MSE
  3. Calculated from means and s's of subgroups
  4. Calculated from a related t or d statistic
  5. Calculated from a related F statistic
  6. Calculated from a related p-value
  7. Standard deviation not used to calculate effect
- *N.* How many subjects were included in the cells involved in your contrast.

- Finally, you must report the effect size, and all calculations you performed to obtain it. This will be covered in more detail in Chapter 5.

4.3 Selecting Moderators

- Sometimes there are differences between the studies that you wish to examine in your synthesis. If you record the important characteristics of each study as variables, you can examine whether the strength of your effect is influenced by these characteristics. 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.
  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 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 test of a moderating variable depends a great deal on the distribution of that variable in your sample. If most of your studies have the same value on a variable, 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 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 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 on 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. As we will discuss in Chapter 6, this violates the assumptions of the standard meta-analysis.
- To reduce the amount of dependence in each analysis, 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 this 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 effect 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.
- 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.
- You might choose to calculate effect sizes from a study that you do not use in any of the analyses. When the effect sizes from your study are based on a common error term you can take a weighted average of those based on a high-level interaction to determine the effect sizes from lower-order effects (see section 5.7). Sometimes this may be easier than directly calculating the lower-order effects.

Chapter 5

Calculating Mean Difference Effect Sizes

5.1 Introduction

- In this chapter we will discuss how to calculate the effect size g and its correction d. For convenience sake we will assume that your contrast will be defined as (experimental group - 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."
- The simplest effect size based on mean differences is Cohen's g, defined as

$$g = \frac{\bar{Y}_e - \bar{Y}_c}{s_p}, \tag{5.1}$$

  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.
- While intuitive, the effect size g is actually a biased estimator of the population effect size

$$\delta = \frac{\mu_e - \mu_c}{\sigma}. \tag{5.2}$$

  Using g produces estimates that are too large, especially with small samples.
- To correct g we multiply it by a correction term

$$J_m = 1 - \frac{3}{4m - 1}, \tag{5.3}$$

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

$$d = g\left(1 - \frac{3}{4m - 1}\right) = g\left(1 - \frac{3}{4(n_e + n_c) - 9}\right) \tag{5.4}$$

  is known as Hedges' d, and is an unbiased estimator of δ. It is generally best to record both g and d for each effect in your meta-analysis.
- The variance of d, given relatively large samples, is

$${\sigma_d}^2 = \frac{n_e + n_c}{n_e n_c} + \frac{d^2}{2(n_e + n_c)}. \tag{5.5}$$

- Using these statistics we can construct a level C confidence interval for δ

$$d \pm z^*(\sigma_d), \tag{5.6}$$

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

- 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 5.4 presented above.

5.2 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

$$g = \frac{\bar{Y}_e - \bar{Y}_c}{s_p}, \tag{5.7}$$

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. The pooled standard deviation can be calculated from the standard deviations of your two groups using the formula

$$s_p = \sqrt{\frac{(n_e - 1){s_e}^2 + (n_c - 1){s_c}^2}{n_e + n_c - 2}}. \tag{5.8}$$

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,

$$g = t\sqrt{\frac{1}{n_e} + \frac{1}{n_c}} = t\sqrt{\frac{n_e + n_c}{n_e n_c}}. \tag{5.9}$$

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

$$g = t\sqrt{\frac{2}{n}} = \frac{2t}{\sqrt{2n}}. \tag{5.10}$$

- From the same logic, if you have a between-subjects z-score comparing the experimental and control groups,

$$g = z\sqrt{\frac{n_e + n_c}{n_e n_c}}. \tag{5.11}$$

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

$$g = \frac{2z}{\sqrt{2n}}. \tag{5.12}$$

- 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),

$$g = \sqrt{\frac{F(n_e + n_c)}{n_e n_c}}. \tag{5.13}$$

If you have the same number of subjects in the experimental and control groups, this equation resolves to

$$g = \sqrt{\frac{2F}{n}}. \tag{5.14}$$

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 5.9 and 5.10. This is because a 1 df F statistic is just the square of a corresponding t statistic.

- 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
  $$L = \sum c_j \bar{Y}_j, \tag{5.15}$$
  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
  $$s_p = \sqrt{\frac{\sum {s_j}^2 {c_j}^2 (n_j - 1)}{\sum {c_j}^2 (n_j - 1)}}, \tag{5.16}$$
  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
  $$g = \frac{L}{s_p}. \tag{5.17}$$

- 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.

5.3 Calculating g Indirectly

- Sometimes a 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. In this case we can reconstruct the appropriate statistic and calculate an effect size.

- 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.

Let us assume two things: you have means for the comparison of interest (say, B_1 versus B_2) but lack a test for this effect. If you have the F and means for another effect (say, A), you can calculate the this effect because you can derive the error term of the ANOVA.

- If there are only two levels of B we can calculate the F statistic associated with this factor and use it to calculate g. To do this we must

  1. Calculate MS_B using the formula
  $$\mathrm{MS}_B = \frac{m_1 + m_2}{4}(\bar{B}_1 - \bar{B}_2)^2, \tag{5.18}$$
  where m_1 is the sample size of the experimental group and m_2 is the sample size of the control group.

2. Calculate MS_E using the formula

$$\text{MS}_E = \frac{\text{MS}_A}{F_A}. \tag{5.19}$$

This is derived from the definition of an F statistic: $F_A = \frac{\text{MS}_A}{\text{MS}_E}$.

MS_A can be calculated using the formula

$$\text{MS}_A = \frac{\sum \left[n_j \left(\bar{A}_j - \bar{G} \right)^2 \right]}{a-1}, \tag{5.20}$$

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

If you are using the F of an interaction (say between A and B) you do the same thing but calculate MS_{AB} using the formula

$$\text{MS}_{AB} = \frac{\sum\sum \left[n_{jk} \left(\overline{AB}_{jk} - \bar{A}_j - \bar{B}_k + \bar{G} \right)^2 \right]}{(a-1)(b-1)}, \tag{5.21}$$

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

3. Calculate F_B using the formula

$$F_B = \frac{\text{MS}_B}{\text{MS}_E}. \tag{5.22}$$

4. Calculate g using formula 5.13.

- If there are ever more than two levels of the treatment factor (even if you have F_B) you calculate s_p from the ANOVA table (equal to $\sqrt{\text{MSE}}$ and use equation 5.7 to determine your effect size. It is not normally possible to calculate an effect in this case if you don't have the means of the experimental and control groups.

- We know that $\sqrt{\text{MSE}}$ is an estimate of the within-cell variance. In a multifactor study, however, this is not σ_p, since it does not include the variance associated with the other factors in the study. To get an estimate of σ_p we need to "reconstitute" the error term by putting back the variance associated with other factors in the model (see Johnson & Eagly, 2000, for a discussion of this issuue). The procedure is to add up the irrelevant sums of squares and pool them with the error sum of squares using the formula

$$s_p = \sqrt{\frac{SS_1 + SS_2 + \cdots + SS_E}{df_1 + df_2 + \cdots + df_E}}. \tag{5.23}$$

You should include all irrelevant main effects, as well as all interactions involving at least one irrelevant factor, in this sum.

This procedure is easy if you have a complete ANOVA table available. If you don't have this table you must reconstruct it yourself. Although difficult, this is possible if you have the cell means and at least one F statistic. You calculate the mean squares of your effects using equations 5.20 and 5.21, and the mean square error using equation 5.19. You can calculate the degrees of freedom directly from your sample size (assuming a balanced design if the marginal counts are not given).

- You have a similar problem if you are only provided with the means and standard deviations on subgroups within the experimental and control groups. The variability associated with the dividing factor has been removed from your standard deviations and must be regained to get s_p. In this case you

1. Calculate $\bar{Y}_e$ and $\bar{Y}_c$ using weighted averages.

2. Calculate the sum of squared scores for the jth subgroup within experimental and control conditions using the equations

$$SE_j = (n_{ej} - 1){s_{ej}}^2 + n_{ej}(\bar{Y}_{ej})^2 \tag{5.24}$$

and

$$SC_j = (n_{cj} - 1){s_{cj}}^2 + n_{cj}(\bar{Y}_{cj})^2, \tag{5.25}$$

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 sum of squares using the formulas

$$(n_e - 1){s_e}^2 = \sum SE_j - n_e(\bar{Y}_e)^2 \tag{5.26}$$

and

$$(n_c - 1){s_c}^2 = \sum SC_j - n_c(\bar{Y}_c)^2. \tag{5.27}$$

4. Plug these two terms into equation 5.8 to calculate s_p.
5. Use $\bar{Y}_e$, $\bar{Y}_c$, and s_p to calculate g from equation 5.7.

- If the irrelevant factors in a study are manipulated (as opposed to being observed), you probably not want to reconstitute their variance into your estimate of s_p. The reason is that the manipulation is actually adding in variance that would not normally be present, artificially reducing the strength of the effect. In this case it would be appropriate to leave this variability out of your estimate of the pooled standard deviation. If you leave this variability out of the error term you might want to code the number of irrelevant variables in the design for each effect size.

- 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

$$s_y = \frac{s_{\text{dif}}}{\sqrt{2(1 - r_{xy})}}, \tag{5.28}$$

where r_{xy} is the correlation between the pretest and posttest scores.

3. Calculate the effect size using the equation

$$g = \frac{\bar{Y}_e - \bar{Y}_c}{s_y} = \frac{\overline{\text{DIF}}_e - \overline{\text{DIF}}_c}{\frac{s_{\text{dif}}}{\sqrt{2(1-r_{xy})}}}, \tag{5.29}$$

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 you do not have this you are probably better off including it as is, even though it will add some error to your analysis.

5.4 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

$$g = \frac{\bar{Y}_e - \bar{Y}_c}{s_{e-c}}. \tag{5.30}$$

If you do not want to work with s_{e-c} you can convert this to the pooled standard deviation using the formula

$$s_p = \frac{s_{e-c}}{\sqrt{2(1-r_{ec})}}, \tag{5.31}$$

where r_{ec} is the correlation between the experimental and control scores. This might prove difficult, however, as few studies report this correlation.

- You can calculate the effect size from within-subjects test statistics using the formulas

$$g = \frac{t}{\sqrt{n}} \tag{5.32}$$

and

$$g = \frac{z}{\sqrt{n}}. \tag{5.33}$$

- 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 term it should be taken. A within-subjects ANOVA has a number of different error terms, and you need to choose the one that would be appropriate to test the contrast in which you are interested. 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 term is obtained you can calculate s_{e-c} using the equation

$$s_{e-c} = \sqrt{2 * \text{MS(within error)}}, \tag{5.34}$$

where MS(within error) is the appropriate within-subjects error term.

- There is a shortcut to figure out which effects are tested using which error terms in a within-subjects design. This can help you determine which error term is used for the treatment factor, which in turn can be used to calculate s_{e-c}. The steps to the shortcut 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 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.

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 5.1: Example output from the shortcut procedure.

So that you can have an idea of how this procedure actually works, figure 5.4 contains the output when it 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.

- 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 completely 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 5.34 to get the standard deviation.

- A within-subjects contrast calculated within levels of a between-subjects variable uses the relevant within 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 d would be calculated from the within-subjects error term using equation 5.34 above.

  If you want to calculate a between-subjects contrast within a within-subjects variable you are dealing with a more complicated situation because such a contrast should use a mixed error term, which is a weighted average of the between error term and the relevant within error term. Therefore, if an effect size for a between contrast is calculated within a within-subjects variable, the standard deviation compiled for the denominator of the d would follow the same logic. Thus, it would average the "between" pooled standard deviation and the "within" standard deviation of differences. If an effect size for a between contrast is calculated within crossed levels of two within variables, all of the standard deviations that are derived from these error terms would be averaged (i.e., the "between" pooled standard deviation and the three relevant "within" 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.

5.5 Estimating g 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 g from the test statistic.

- You can get inverse probability distributions from a number of statistical software packages, including 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 g), 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 $g = 0$, 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.

5.6 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 chapter 17 of the Handbook.
- One method, 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
  $$g = (z_e - z_c)\sqrt{\frac{n_e + n_c}{n_e n_c}}. \tag{5.35}$$
  When you have the same number of subjects in the experimental and control group this equation resolves to
  $$g = \frac{2(z_e - z_c)}{\sqrt{2n}}. \tag{5.36}$$
- 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
  $$\bar{Y} = p \tag{5.37}$$

and

$$s = \sqrt{pq}, \tag{5.38}$$

where q is defined as $1 - p$.

4. Calculate the pooled standard deviation, using equation 5.8. This equation becomes

$$s_p = \sqrt{\frac{(n_e - 1)p_e q_e + (n_c - 1)p_c q_c}{n_e + n_c - 2}} \tag{5.39}$$

5. Use $\bar{Y}_e$, $\bar{Y}_c$, and s_p to calculate g using equation 5.7.

- 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 do have the frequencies you should use one of the two methods presented above.
  - 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

$$g = \sqrt{\frac{\chi^2(n_e + n_c)}{n_e n_c}}. \tag{5.40}$$

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

$$g = \sqrt{\frac{2\chi^2}{n}}. \tag{5.41}$$

    You can alternatively calculate the phi coefficient using the equation

$$r_\phi = \sqrt{\frac{\chi^2}{n}} \tag{5.42}$$

    and calculate g from r using equation 5.44.
  - If one or both levels has more than 2 levels, you can calculate

$$P = \sqrt{\frac{\chi^2}{n + \chi^2}}, \tag{5.43}$$

    which approximates r if the sample size is large. You can then transform r to g using equation 5.44.

5.7 Calculating g by Averaging Other Effects

- You can calculate the size for a effect averaging over the levels of a variable by combining the g's from the different levels of the averaged variable, if all of the effects under consideration use the same error term. For example, in a meta-analysis examining the impact of target gender on evaluations, let us assume that you are calculating three effects from a single study: one which is the effect for male subjects, one which is the effect for female subjects, and one with is the overall effect for the study (averaging over gender). Since all of the effects in the study are based on the between-subjects error term you could calculate the g of the overall effect by taking the mean of the g's from the other two effects.
- You cannot calculate d this same way because the transformation from g to d is dependent on the number of subjects composing the effect which will typically not be the same when combining effects together. However, it is relatively simple to recalculate d using equation 5.4.
- If the study of interest uses a within-subjects design you should be very careful because an effect that averages over a within-subjects factor will *always* use a different error term. Always double check to make sure that your effects use the same error term before applying this method.

5.8 Miscellaneous

- To calculate g from r you use the formula

$$g = \frac{2r}{\sqrt{1 - r^2}}. \tag{5.44}$$

- 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 5.5). You then calculate g using equation 5.9. For more precision you can make an adjustment for the lower power of the nonparametric statistic (see Glass, McGaw, & Smith, 1981).
- You should always report g and d statistics to four decimal places.

Chapter 6

Calculating Correlation Effect Sizes

6.1 Introduction

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

$$r_{xy} = \frac{\sum z_{xi} z_{yi}}{n}, \tag{6.1}$$

  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 relationship, correlations near 1 indicate a strong positive relationship, while correlations near 0 indicate no linear relationship.

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

$$G_{(r)} = r + \frac{r(1-r^2)}{2(n-3)}. \tag{6.2}$$

- 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

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \tag{6.3}$$

  where $\ln(x)$ is the natural logarithm function. All meta-analytic calculations are then performed using the transformed values.

- 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-scores for analysis.

- z_r has a nearly normal distribution with variance

$$s_z{}^2 = \frac{1}{n-3}. \tag{6.4}$$

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

$$z_r \pm \frac{z^*}{\sqrt{n-3}}, \tag{6.5}$$

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

- Once you have made the necessary computations, you use Fisher's z-to-r transformation

$$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}, \tag{6.6}$$

where e is the base of the natural logarithm (approximately 2.71828), to convert the results (ex., mean effect size, confidence interval boundaries) back into correlations.

- Meta-analysts have also 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 6.2, and then convert this to a z-score using formula 6.3 before analyzing the effect sizes.

6.2 Calculating r from Linear Regression

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

$$y = b_0 + b_1 x_1 \tag{6.7}$$

and x_1 and y are your two variables of interest you can calculate $r_{y,x1}$ using the equation

$$r_{y,x1} = b_1 \left(\frac{s_{x1}}{s_y} \right), \tag{6.8}$$

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

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n, \tag{6.9}$$

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 modeling, including Kenny (1979, p. 31-33).

6.3 Calculating r from Test Statistics

- As mentioned above, the correlation coefficient is designed to measure the linear relationship 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 relationship 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 relationship 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 relationship 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

$$r_b = \frac{r_{pb}\sqrt{n_e n_c}}{|z^*|(n_e + n_c)}, \tag{6.10}$$

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 relationship between two truly dichotomous variables. This actually is an r.

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

$$r_{pb} = \sqrt{\frac{t^2}{t^2 + n_e + n_c - 2}}. \tag{6.11}$$

  You can then transform r_{pb} into r_b using equation 6.10 to get an estimate of r.

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

$$r_{pb} = \sqrt{\frac{F}{F + n_e + n_c - 2}}. \tag{6.12}$$

  You can then transform r_{pb} into r_b using equation 6.10 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 relationship between your variables. See section 5.2 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

$$r_{\cos-\pi} = \cos\left(\frac{180^\circ}{1 + \sqrt{\frac{ad}{bc}}}\right). \tag{6.13}$$

- 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

$$r_\phi = \sqrt{\frac{\chi^2}{n}}. \tag{6.14}$$

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

$$r_{pb} = 1 - \frac{2U}{n_e n_c}, \tag{6.15}$$

  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 6.10.

6.4 Miscellaneous

- You can calculate r from g using the equation

$$r = \sqrt{\frac{g^2 n_e n_c}{g^2 n_e n_c + (n_e + n_c)(n_e + n_c - 2)}}. \tag{6.16}$$

- You can calculate r from d using the equation

$$r = \sqrt{\frac{d^2}{d^2 + 4}}, \tag{6.17}$$

  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

$$r = \sqrt{\frac{d^2}{d^2 + \frac{1}{pq}}}, \tag{6.18}$$

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

- You should always report r to 4 decimal places.

Chapter 7

Issues in Calculating Effect Sizes

7.1 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 formulas, however, are not all equally valid – some involve making more inferences than others. In general, you want to calculate your effect size as directly as possible. The more steps you have to take, 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 between-subjects t test for the effect of interest
  - a within-subjects 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 d
  - $\bar{Y}_e$, $\bar{Y}_c$, 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 sometimes directly reported. The best methods are to calculate r from
  - a directly reported correlation

- simple linear regression

The second class of methods calculate r from

- multiple regression
- 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 relationship 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.

7.2 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.
- 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.
- Researchers have developed ways to correct effect sizes for measurement unreliability, restriction of range, artificial dichotomization of variables, and imperfections in construct validity. These are detailed in Chapter 21 of the Handbook.

7.3 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 g statistics you can use the formula

$$\text{combined } g = \frac{\sum g_i}{\sqrt{\rho m^2 + (1-\rho)m}}, \tag{7.1}$$

  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.

- To combine several correlations you can use the formula

$$\text{combined } z_r = \frac{\sum z_{ri}}{\sqrt{\rho m^2 + (1-\rho)m}}, \tag{7.2}$$

  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

- 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.

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

$$\bar{r}_{ij} = \frac{\alpha}{n + (1-n)\alpha}, \tag{7.3}$$

  where α is Chronbach's alpha and n is the number of response measures. r_{ij} can be used for ρ in equations 7.1 and 7.2.

7.4 Combining studies from different designs

- As discussed briefly in section 5.4, effect sizes calculated from between-subjects designs are based on the standard deviation pooled across groups, s_p, while those calculated from within-subjects designs are based on the standard deviation of the difference score, s_{e-c}. This means that effect sizes calculated using the two different designs are not directly comparable and should not both be used in the same analysis.

- The best solution to this problem is to convert all of your effect sizes so that they are all based on the same variance estimate. If you have a d calculated based on the standard error of the difference score (such as from a within-subjects design) you can calculate an equivalent d based on the pooled standard error from the formula

$$d_p = d_{e-c}\sqrt{2(1-\rho)}, \tag{7.4}$$

  where ρ is the correlation between the experimental and control group in the within-subject condition. Similarly, you can calculate a d based on the standard error of the difference score from one based on the pooled standard error using the formula

$$d_{e-c} = \frac{d_p}{\sqrt{2(1-\rho)}}, \tag{7.5}$$

  where ρ is again the correlation between the experimental and control scores in the within-subject condition.

- Unfortunately, very few studies will provide sufficient information to calculate ρ, especially when the original d was derived from a between-subjects design. What you may choose to do is to calculate an average ρ from the studies that do report sufficient information and then use this for studies for which ρ is unknown.

  For more information about combining effect sizes calculated from different designs see Morris and DeShon (2002).

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.
- 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 should 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 should 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

- 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.

- You should also report the median and unweighted mean effect sizes of your sample to give your audience a better sense of the central tendency. 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.

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

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

- 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 homogeneity 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.

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

  1. Calculate the homogeneity 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 homogeneity.
  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.

8.4 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.
- 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	d	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

$$X = \frac{(\sum z_i)^2}{2.706} - N_L, \tag{8.7}$$

  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 present a 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

$$\text{success}_E = .5 + \frac{r}{2} \tag{8.8}$$

3. Calculate the success rate of your control group

$$\text{success}_C = .5 - \frac{r}{2} \tag{8.9}$$

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. 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.

8.5 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 is drawn directly from the cumulative 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 (including 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

Examining Moderating Variables

9.1 Relationships Between Moderators

- Very rarely will you find that the moderators in your study are crossed with each other. 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 relationship between two variables (such as a moderator and the effect size) you need to not only show that a relationship exists between the two but also that the relationship 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 relationships 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 examine the relationship 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 relationship between your two moderators.
- The easiest way to examine the relationship between two continuous moderators is to calculate the Pearson correlation. If you want to test for a more complicated relationship (such as quadratic) you can use regression analysis and test the values of the coefficients.
- To examine the relationship 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 relationship using ANOVA.
- You might consider weighing each study in these analyses by the sample size. However, you also 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 for a large number of characteristics, it is usually not feasible to test for covariation between every pair of moderators. You will likely be better off if you select a subset of the possible relationships to test based on theoretical considerations.

- When presenting the relationships between moderators in your analysis, you should choose a single statistic (such as correlations) for presentation purposes. You should convert all measures of association that you calculate into the chosen form. This will make it much easier for your readers to interpret the results.

9.2 Models Predicting Effect Sizes

- The ultimate goal of most meta-analyses is to predict variations in effect sizes from the values of moderating variables. This procedure requires that the analyst (at least implicitly) specifies a theoretical model.
- The two most general categories of models are *fixed effect* and *random effect* models.
  - Fixed effect models allow 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.
  - Random effect models assume that your the studies you observed are a random sample, allowing you to generalize to the population from which the sample was drawn. This allows you a great deal more freedom to apply your inferences to new situations.

  While random effect models may sometimes be more appropriate, almost all meta-analyses are conducted using fixed effect models because the mathematics behind them are much simpler. The methods that we will consider here are all based on fixed effect models.
- 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.
- We will consider two general classes of fixed effect models, the first paralleling one-way ANOVA and the other paralleling simple linear regression. After this we will discuss how you can use an extension of regression to test larger meta-analytic models.

9.3 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 homogeneity 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 can break Q_T into two parts: the variability that can be explained by a moderator and the variability that cannot.
- We use the between-groups homogeneity Q_B to measure how much variability can be explained by a moderator. To calculate Q_B 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

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

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

$$w_j = \frac{1}{s_{\bar{T}j}{}^2}, \tag{9.2}$$

  and the summation is over the different 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 homogeneity Q_W to measure how much variability the moderator fails to explain. To calculate Q_W you

  1. Calculate the variability within each individual level of your moderator. The variability Q_{Wj} for level j is simply the homogeneity (see equation 8.6) of the studies within level j.
  2. Calculate Q_W using the equation
  $$Q_W = \sum Q_{Wj}, \tag{9.3}$$
  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 homogeneity. 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. The first thing you need to decide is whether you will perform *a priori* or *post hoc* contrasts. A priori contrasts are appropriate when you want to compare two groups based on some theoretical issue you considered before performing your moderator analyses. Post hoc contrasts are appropriate when you want to figure out which of a group of means are different as a follow-up to a significant moderator test.

  - To perform an a priori contrast you
    1. Define your contrast as a linear combination of the group mean effect sizes, using the form
    $$L = c_1\bar{T}_1 + c_2\bar{T}_2 + \cdots + c_p\bar{T}_p, \tag{9.5}$$
    where p is the number of levels in your moderator and $c_1 + c_2 + \cdots + c_p = 0$.
    2. Calculate the value of the contrast. This will provide you with information about the direction of the results.
    3. Calculate the statistic
    $$\chi^2 = \frac{L^2}{{c_1}^2\bar{T}_1 + {c_2}^2\bar{T}_2 + \cdots + {c_p}^2\bar{T}_p}, \tag{9.6}$$
    which follows a chi-square distribution with 1 df. Large values of χ^2 indicate that the value of your contrast is significantly different from zero.
  - To perform a post-hoc contrast you use the same procedure, but you must adjust the distribution of the χ^2 statistic to account for the total number of post-hoc contrasts you perform. According to the Scheffe method you compare χ^2 to the chi-square distribution with B degrees of freedom, where B is equal to the smaller of (the number of contrasts you perform) and $(p-1)$. According to the Bonferonni method you use the chi-square distribution with 1 df, but you divide the p-value you will accept as significant by the total number of post-hoc contrasts you perform. The Bonferonni method is more conservative than the Scheffe method.

9.4 Testing a Continuous Moderator

- You can test whether there is a linear relationship 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 chapter 19 of the Handbook.

- 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 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 relationship between effect size and your moderator.

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 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 9.1 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 The 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.
  - 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 The Method Section

- 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 begin reviewed.
    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.
  - 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 The Results Section

- In the results section you describe the distribution of your effect sizes and present any moderator analyses you decided to perform.
- 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 homogeneity 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 homogeneity Q_{Wj} and its corresponding p-value.
  - The between-group homogeneity Q_B and its corresponding p-value.
  - The total within-group homogeneity 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 ability of a statistical model to explain your distribution of effect sizes you should present
  - 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.
  - Any specific parameter tests you wish to discuss.

10.5 The Discussion and Conclusion

- 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 8.4 designed to provide intuitive meaning to effect sizes.
- You should attempt to provide an explanation for any significant moderators revealed by your analyses.
- 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.

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 is aggregating. 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. For most purposes you would want to have at least 30 studies.
- If a meta-analysis performs moderator tests it should also report if there are any relationships 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.
- 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. 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 in their computer searches, and what methods they used to locate studies other than computer searches.
  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?

  Remember, 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.

- Chapter 29 of the handbook (by Miller & Pollock) 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. (1977). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Academic Press.

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.

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.

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.

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

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.

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.