CHAPTER 36

# Meta-Analysis in Personality Psychology

## *A Primer*

**Brent W. Roberts**
**Nathan R. Kuncel**
**Wolfgang Viechtbauer**
**Tim Bogg**

We assume that if you are reading this chapter, you are interested in doing a meta-analysis. Good for you. We believe more researchers should take advantage of this technique. In fact, if you have ever done a literature review, then you probably have examined enough studies to conduct a meta-analysis. Moreover, analytically speaking, meta-analyses are not unusually difficult. The fact that two of us (B. W. R. and T. B.) have published meta-analyses is testament to the fact that the technique can be mastered by the mathematically challenged. That is not to say that they are easy to do. A good meta-analysis usually takes more time and effort than a typical study, despite the fact that you seldom collect your own data. Just don't let the analytical barriers faze you.

In this chapter we provide an overview of the steps one should take in performing a meta-analysis. Our treatment is by no means exhaustive and does not replace one of the numerous in-depth descriptions of the technique (see "Recommended Readings"; Cooper & Hedges, 1994; Hedges & Olkin, 1985; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001; Rosenthal, 1991). Our goal is to provide the reader with a decipherable overview of the steps taken in most meta-analyses within personality psychology. In addition, we point out more authoritative sources for the specifics of meta-analysis. The reader should be forewarned: There will be a few formulas along the way, but these are no more complicated than those found in elementary discussions of classical test theory, and when examined closely, can actually illuminate several key issues.

The chapter is organized around the steps typically taken in conceptualizing and performing a meta-analysis. In the first section we discuss why meta-analysis is useful. The second section deals with aspects of data collection. Inasmuch as the data for a meta-analysis are derived from a set of related studies, issues in conducting a thorough literature review are

considered in some detail. The third section focuses on a critical stage of the meta-analysis—organizing and coding the studies before analyzing them. The fourth section describes the varied approaches to analyzing meta-analytic data. Finally, the fifth section touches on some of the issues that we have been confronted with in our own meta-analytic work and that may be particularly germane to personality psychologists.

## Why Do a Meta-Analysis?

Imagine the following scenario. An investigator is interested in linking the trait of conscientiousness to tobacco consumption (e.g., Bogg & Roberts, 2004). Across three studies she finds correlations of –0.33, –0.15, and –0.25, suggesting that higher levels of conscientiousness are related to less tobacco consumption. However, with modest sample sizes of 75 participants in each study, one would find only the –.33 correlation to be statistically significant. What should the researcher conclude on the basis of these findings? Should the three studies be written up as one rejection of the null hypothesis and two failures to reject? If this approach is taken, how likely is this researcher's article to be published? Given our reliance on null hypothesis significance testing for determining the existence of an effect, the reviewers would inevitably conclude that the original study failed to replicate, twice.

This scenario is all too common. One can find countless examples of researchers concluding that either their newly collected data or already published findings are contradictory because some effects are statistically significant whereas others are not. We often rely on this apparently contradictory pattern to justify new research, as we believe that through some ingenious methodological innovation of our own we will be able to rectify the discrepancy. Unfortunately, the conclusion that the results are contradictory is often erroneous for one specific reason. Most of our studies are woefully underpowered (Cohen, 1992). For better or worse, personality psychologists tend to study small groups (e.g., 50–150 participants). Given the modal effect size of our research and the low power of studies in personality psychology, we often have only a 50:50 chance of determining that a small or medium-size effect is statistically significant. Such odds are not particularly comforting. The tragedy is that we continue to design studies in this way despite knowing better (see Fraley, Chapter 8, this volume).

This problem highlights one of the first and most fundamental reasons to do a meta-analysis—to ask the question, "Is there an effect?" or, more accurately, "What is the magnitude of the effect?" given the fact that the effect is seldom "nill." In the case of our example, when combined meta-analytically, the effect size is –.24, and voilà, it is statistically significant. Is this a small effect? Not in terms of the normal range of effect sizes found across psychology and medicine (Meyer et al., 2001). Thus, it would have been an inferential error to conclude that there was no effect. The preponderance of underpowered studies in personality psychology alone is sufficient justification for combining the results from several commensurable studies with meta-analytic methods. And, as noted above, rather than conducting a narrative review of the literature that comes to the inevitable conclusion that the research is contradictory, why not run a small meta-analysis and derive a point estimate that might frame the results of your study more concretely and precisely?

Alternatively, if a domain has a rich history, then one can perform a more exhaustive meta-analysis in order to determine whether an effect exists and how large it is. This is exactly what we did with our meta-analysis of the relationship between conscientiousness and health behaviors (Bogg & Roberts, 2004). We were optimistic that conscientiousness would be related to at least some of the leading behavioral contributors to premature mortality. What we were truly interested in was whether it was related to *all* of the behaviors, which it was. Thus, we did a relatively straightforward "is there an effect here" meta-analysis that resulted in profoundly important results—conscientiousness has pervasive relationships to all of the reasons people die prematurely.

The second reason to do a meta-analysis is exactly the same reason we do any study—to test hypotheses or compare models. Just like any other analytical technique, such as regression, analysis of variance, or even the simple correlation coefficient, meta-analysis can be used to test specific hypotheses derived from different theories. For example, in our two meta-analyses of longitudinal personality trait development, we tested the theory that personality traits stop changing after age 30 and

found convincing evidence leading us to reject the "no change after 30" hypothesis in both cases (Roberts & DelVecchio, 2000; Roberts, Walton, & Viechtbauer, 2006a). Here, we used meta-analysis like any other statistical technique—as a hypothesis-testing tool.

Once you have decided to conduct a meta-analysis, what are the basic steps and issues to consider at each of these stages? If the devil is in the details, then conducting a meta-analysis is surely fraught with danger. Executing a large-scale meta-analysis involves considerable management of information to minimize wasted effort, ensure precision, and avoiding the agony of needing to redo parts of the study. Although work on each of these phases may overlap, it is valuable to think of a meta-analysis as being broken into five key phases: Conceptualizing the problem to be studied, identifying and obtaining articles, coding and proofing the data, and analyzing data and reporting the results. Table 36.1 provides a checklist of issues to consider at each of these stages. This is not a hard-and-fast checklist, as several decisions made along the way will affect which steps are actually taken. Nonetheless, we believe that most of the important and pragmatic issues are highlighted.

**TABLE 36.1. Meta-Analysis Checklist**

I. *Conceptualizing the problem*
   1. Research question/hypothesis
   2. Level of analysis

II. *Identifying and collecting articles*
   1. Search databases and journals
      a. PsychLit, PubMed, etc.
      b. Conference proceedings and programs
      c. Technical reports
      d. Relevant journals
      e. Review articles
      f. Dissertations
   2. Search out fugitive literature
   3. Snowballing
      a. Search references in articles in the database
      b. Citation index search of all articles

III. *Coding articles*
   1. Create coding protocol
   2. Coder training
   3. Coding and periodic coding checks

IV. *Preparing the data*
   1. Transforming effect sizes
   2. Directionalizing effect sizes
   3. Aggregating nonindependent effect sizes
   4. Consider correcting for artifacts

V. *Analyzing your data and reporting results*
   1. Choose a model: fixed effects, random effects, mixed effects
   2. Test for publication bias
   3. Test for moderators
   4. Aggregating effect sizes and reporting your results

## Conceptualizing the Problem to Be Studied

It is critical to clearly specify the research question to be answered before the literature search is actually conducted. Doing so will be invaluable for the same reasons this step is critical for primary research. Most important, having a clear conceptualization of the primary research questions helps one to make better decisions with respect to the relevant information one should gather. This, in turn, will focus the literature search and ensure that the project does not spin out of control.

As in any study, the primary issue is the tradeoff between breadth and specificity. If the research question is too broad, the resulting meta-analysis may be too diffuse to answer the original research question well. For example, meta-analytic research is often criticized for mixing "apples and oranges" because the studies being aggregated are different from one another in some key way. Of course, if you are interested in fruit, then the broader level of conceptualization is appropriate. Conversely, if the research question is too narrow, there may be too few studies to analyze and the question may be of little interest to all but a few readers. Striking the right balance between breadth and specificity by an appropriately formulated research question is one of the most critical issues when doing a meta-analysis.

The key arbiter of getting the tradeoff right is experience. The entire meta-analytic enterprise, and especially the conceptualization of the research question, will be much easier for those who have long toiled in the back alleys of primary data collection and know a research area well. Approaching a new area for meta-analytic investigation with little experience is fraught with numerous conceptual and procedural potholes. The idea may not be sound. The key studies may escape the search. A meta-analysis on the topic may already have been

published. These issues should not stop a new researcher from doing a meta-analysis. What it should do is motivate the researcher to do his or her homework and to talk the issues through with people actively involved in the particular area of research. Once a clear plan for a meta-analysis has been acquired through hard work or high-quality consultation, you are ready to start identifying and collecting articles (i.e., "data collection").

## Identifying and Collecting Articles

Given the advent of the personal computer and the ever improving databases and search engines, one might conclude that identifying the relevant articles is as easy as doing a PsychLit search. This is not the case. A thorough and valid search of the literature encompasses using electronic databases and journals, searching for the fugitive literature, and then "snowballing" existing articles. We describe each of these techniques in turn.

When it comes to electronic database searches, we recommend the following procedure. First, generate a list of key words and conduct searches with the key words used in various combinations. It is critical to remember that many social science domains are studied by more than one discipline. Conducting searches using the databases and jargon of these fields is valuable for creating a thorough search. If the resulting lists are too large, restrict key words to the title of the paper. In general, it is far better to conduct a too broad search rather than one that is too narrow. In addition, detailed notes regarding the key words used and the steps followed during the search process will be invaluable for rerunning searches at a later point.

All searches should then be imported into a bibliographic database (e.g., Endnote, Reference Manager, Refworks). These searches can be supplemented by searching conference programs, conference proceedings, and technical report lists for relevant organizations. Many bibliographic database programs allow for the identification of identical references. After duplicates are deleted, the database can be further edited by examining each abstract to make a judgment about its relevance. References *should not* be deleted; rather, they should be labeled as rejected, using another field in the bibliographic database manager.

Data for some topics can be found in journal articles without being the central focus of the study. This makes electronic searches challenging. For example, in a meta-analysis on the validity of self-reported grade point averages (Kuncel, Credé, & Thomas, 2005) it became apparent that the pertinent data were often mentioned in a method section about the outcomes measured without being the actual focus of the study. A combination of approaches can be used to address this common problem. The first is to identify specific topics that seem to frequently contain the desired information even if it is not the central focus of the study. Such ancillary topics can then be searched using typical methods described above. Second, one can identify journals that seem to most commonly contain the necessary information and conduct hand searches of those journals.[1] Third, reading review articles, like those found in the *Annual Review of Psychology* or in the constantly proliferating handbooks now being published every week may reveal articles that are not found in a typical electronic database search. Note that these techniques could be used for any meta-analytic topic to enhance the comprehensiveness of the search.

Once a database of desired articles has been constructed, the items need to be collected. The bibliographic database software can be used to generate a compact list of articles to be collected. It is important at this stage to be kind to your interlibrary loan personnel as they can make this stage far easier. Fortunately, many documents are now available in electronic format and can be directly downloaded, including dissertations and technical reports. Key researchers that show up frequently in the database can also be contacted at this stage to see if they have other published studies that were overlooked during the literature search or unpublished studies they would be willing to share.

As the articles enter the laboratory, the bibliographic database needs to be updated to note that the articles are now "in house." A filing system can be created that facilitates the processing of articles. At a minimum, there should be space for new articles, articles that have been coded, articles that have been processed (data and "snowballs" entered), and articles that have been proofed and are ready for more long-term storage.

The next to last step in the search is concerned with the "fugitive literature." This step

in the search process is especially important in that it may result in the inclusion of a number of studies that report null effects, as the field is nearly uniformly biased against publishing null findings.[2] The inclusion of these studies should provide more accurate estimates of effect sizes. These studies can be discovered through requests to list serves. Another technique we have found useful is to send the initial list of studies included in the meta-analysis to key individuals who have studied the phenomena of interest. These researchers can often identify unpublished studies and studies that have unusual titles that do not show up in the typical search procedures.

The last step taken, once the initial database has been compiled, is to snowball the preliminary database. First, the references of articles included in the meta-analysis or the references found in review articles should be examined for studies that were missed during the electronic search. Second, papers that reference the studies in the meta-analytic database should be examined to see if they report similar data (i.e., a citation index search). Review articles and dissertations, owing to their lengthy review sections, are especially helpful for snowballing. The easiest approach is to simply mark in the reference lists of the articles that look promising. To avoid duplicate collection of articles, these noted references should be compared against the updated bibliographic database to see if an article has already been identified and collected. New articles can be added and flagged using a separate field. Depending on the topic, snowballs can easily increase a database by 30–50%. Efficiently collecting this information is dependent on having a well-managed bibliographic database.

As indicated by the GIGO acronym used to deride factor analysis (i.e., garbage in, garbage out), a meta-analysis is only as good as the studies it examines. In addition to being one of the most critical stages of the process, literature identification will also take much longer than typically expected. Its importance should not be underestimated. To use a sports metaphor, it is like getting the footwork right for a tennis shot. If the feet are not in the right place, it matters little how well the person swings; the player will still miss the shot. Don't make your meta-analysis a swing and a miss. Put the necessary time and effort into the data collection stage, and you will be rewarded with a definitive study.

## Coding Articles

Once a suitable body of literature has been identified and collected, the next task is to extract information from each study that will be used in the subsequent analyses. The most common means of extracting information from research reports and other data sources is a coding protocol. A coding protocol (sometimes initially guided by a coding manual) is a form used by coders to document two distinct types of information from data sources: (1) study descriptors—information regarding the characteristics of the study, also called *moderators*; and (2) effect sizes—information regarding the actual findings of the study (Lipsey & Wilson, 2001).

The key to the successful development of any coding protocol is planning. Decisions need to be made early on regarding study descriptor and effect size information that is relevant for the meta-analysis. Some of these decisions will be guided by the a priori investigative goals of the meta-analyst (e.g., gender is expected to moderate the effect of interest). Other decisions will require a review of the collected body of literature (or a representative subsample thereof) to determine which study descriptor and effect size information occurs with sufficient frequency to warrant inclusion in the coding protocol. Even if a particular study descriptor (e.g., ethnicity) is of interest, a review of the collected literature may reveal it to be reported so infrequently that requiring coders to document it across studies would be unproductive.

A list of potential moderators and outcomes should be identified at the beginning of the project. We should note that the term *moderator* is analogous to *independent variable* in primary data collection. In meta-analytic jargon this reflects the fact that an independent variable that is related to variability in meta-analytic outcomes is directly analogous to a moderator effect in a typical study. As coding proceeds, it is often the case that new variables may appear that are interesting. For example, a new outcome may appear in a few studies that had not been considered before. New fields or codes should be created for these variables in the coding sheet, and previous studies should be reexamined. However, it is important to note that moderator analyses can be overdone. It is valuable to think of a database as having a limited amount of information value. A vast number of thoughtless moderator tests can, by chance,

yield an apparently important moderator. It is best to avoid this shotgun approach to research.

A moderator that is common to many meta-analyses is the "study quality" moderator, in which the researcher makes a global evaluation of whether a study is of high quality or not. In practice, this is often a subjective judgment made by those coding the studies. This approach can easily fall prey to the coder's biases regarding theories, journals, methods, and even other scientists. It is our position that such a subjective approach should be avoided. Given the vast number of books and articles on experimental, correlational, and quasi-experimental design, study quality can be thoughtfully and specifically operationalized. That is, as scientists we should be able to clearly specify how and why one study is of lower quality than another. In many cases, multiple study characteristic codes will be necessary to capture this information.

As mentioned above, the types of study descriptors coded for each study are dependent on the declarative and exploratory interests of the meta-analyst. At the broadest level is information about the source of the study (i.e., journal, dissertation, book, etc.) and its year of publication. Information about the study's author(s) may also be of interest, as well as any sources of funding for the research (Lipsey & Wilson, 2001). Of more substantive interest are study characteristics that have a direct or an indirect bearing on the relationship being investigated. These characteristics include the source of the sample (e.g., a long-standing national or regional study), demographic information about the sample (e.g., gender, age, ethnicity, socioeconomic status), and other identifying features of the sample (e.g., clinical versus nonclinical, inpatient versus outpatient, delinquent, criminal).

Perhaps the most important study descriptors are those related to the independent and dependent variables. These characteristics include the types of independent and dependent variables employed (usually described in terms of constructs and their forms of operationalization) and the quality of the measures used (e.g., reliability). For example, in a meta-analysis investigating the relationship between extraversion (independent variable) and exercise (dependent variable), the coding protocol would provide options for which construct related to extraversion was investigated in the study (e.g., extraversion, social dominance, sociability, activity) as well as how it was measured (e.g., NEO-FFI, California Psychological Inventory). Similarly, options for specifying the exercise-related construct (e.g., strength, flexibility, endurance, cardiorespiratory fitness) as well as the means of measurement (e.g., maximal bench press, $VO_2$ maximal treadmill test) would be provided. In this way, the coding protocol behaves as a survey, providing the coder with response options or the ability to provide an "open" response.

In terms of the actual statistical analyses of the studies, effect size information must be carefully considered and coded. At the very least, there are two statistics that must be entered into the coding protocol for each study—the effect size statistic and the sample size specific to that effect size. This information is crucial for meta-analytic calculations. There are also other features—some statistical, some conceptual—that are desirable to code. Additional effect size information includes a description of the variables that comprise the effect size (described by construct labels, measures, or both), subsample information (relevant when multiple effect sizes are coded across different configurations of a sample or multiple samples in a study), standard deviations, reliability of variables comprising the effect size, dichotomization of variables comprising the effect size, statistical transformation procedure (how an effect size was calculated if the desired metric was not available in the study, e.g., using means and standard deviations to calculate a correlation coefficient), a confidence rating for the effect size (coder-rated level of surety in the integrity of the coded effect, usually lower for crude estimations), and a page number or other location information (e.g., table) where the effect size information (or any other characteristic of the study) can be double-checked for accuracy. As with the study descriptors, decisions regarding the inclusion of effect size information should be made based on an understanding of which information is desired and typically available in the collected body of studies.

As effect sizes are so important to the meta-analytic approach, we now discuss in detail some effect size measures frequently found in the personality research literature, namely the standardized mean difference for two independent groups, the standardized mean difference for two dependent groups, the raw product–moment correlation coefficient, and the corre-

lation coefficient after applying Fisher's variance stabilizing transformation. It should be noted that the effect size measures discussed are just a selection of a large number of effect size indices that can be calculated. They were chosen for a more detailed description because of their ubiquitous use in personality research and for illustrative purposes, but not as an argument for their superiority to other effect size indices. The choice of an effect size measure is partly dependent on the types of studies being meta-analyzed and on the reporting practices within a research community. Because the types of studies and reporting practices can differ widely, a large variety of effect size indices are available and have been described in detail in the existing literature (e.g., Fleiss, 1994; Lipsey & Wilson, 2001; Rosenthal, 1994).

## Standardized Mean Difference for Two Independent Groups

The standardized mean difference (SMD) measures the mean difference between two independent groups on some continuous outcome measure, in which one group can be considered the experimental (E) and one the control group (C). Because the raw units of outcome measures across studies are typically not commensurable (e.g., a 5-point mean difference between two groups on the California Psychological Inventory (CPI) Dominance scale may reflect a larger/smaller difference than a 5-point difference on the Jackson Personality Inventory (JPI) Dominance scale), we must first find a way to make different scales comparable across the studies. This can be accomplished by dividing (standardizing) the raw mean difference by the pooled standard deviation of the two groups.

Therefore, assume that for each study, the scores within the two groups are normally distributed with means $\mu^E_i$ and $\mu^C_i$ and common variance $\sigma^2_i$. Then the effect size in the $i$th study is given by

$$\theta_i = \frac{\mu^E_i - \mu^C_i}{\sigma_i}$$

which we can estimate with

$$d_i = \frac{\bar{x}^E_i - \bar{x}^C_i}{s^P_i}$$

where $\bar{x}^E_i$ and $\bar{x}^C_i$ are the observed sample means and $s^p_i$ is the pooled standard deviation of the two groups. However, $d_i$ tends to be slightly too large on average (it overestimates $\theta_i$). One can easily correct this bias by computing

$$ES_i = \left(1 - \frac{3}{4m_i - 1}\right) d_i$$

where $m_i = n^E_i + n^C_i - 2$ and $n^E_i$ and $n^C_i$ are the sample sizes of the two experimental groups (Hedges, 1981).

The sampling variance of $ES_i$ can be calculated with

$$vi = \frac{n^E_i + n^C_i}{n^E_i n^C_i} + \frac{ES^2_i}{2(n^E_i + n^C_i)}$$

Therefore, $v_i$ denotes the amount of variance expected in the effect size estimate due to sampling fluctuations alone. As the sample sizes ($n^E_i$ and $n^C_i$) of the two experimental groups increase, $v_i$ becomes smaller, reflecting the fact that effect size estimates based on larger samples tend to be closer to their corresponding $\theta_i$ value.

## Standardized Mean Difference for Dependent Samples

The SMD can also be used when the same group of subjects is measured at two points in time, such as before and after receiving some kind of treatment or as part of a longitudinal study to examine changes across time. Because the same group of subjects is measured twice, the subjects' scores can no longer be assumed to be independent. Specifically, when $j = 1, \ldots, n_i$ subjects are tested at two time points, T1 and T2, and the scores at the two time points are normally distributed with means $\mu^{T1}_i$ and $\mu^{T2}_i$ and common variance $\sigma^2_i$, then we expect there to be a certain amount of correlation between the scores at T1 and T2, which we denote with $\rho_i$.

The raw change across time, given by $\mu^{T2}_i - \mu^{T1}_i$, is typically not a useful effect size measure in meta-analysis, because the units of the various outcome measures across the studies are not directly comparable. The solution again is to standardize the raw mean difference in some way, and two options for doing so have been suggested in the literature (Morris & DeShon, 2002).

### *Raw Score Metric*

The first option is to standardize the mean change by the standard deviation of the raw scores, yielding the effect size

$$\theta_i = \frac{\mu_i^{T2} - \mu_i^{T1}}{\sigma_i}$$

An estimate of $\theta_i$ is given by

$$ES_i = \left(1 - \frac{3}{4m_i - 1}\right)\frac{\bar{x}_i^{T2} - \bar{x}_i^{T1}}{s_i^{T1}}$$

where $m_i = n_i - 1$, $\bar{x}^{T1}{}_i$ and $\bar{x}^{T2}{}_i$ are the observed sample means at the two time points, and $s^{T1}{}_i$ is the observed standard deviation of the scores at time T1. The sampling variance of $ES_i$ can be calculated with

$$v_i = \frac{2(1 - r_i)}{n_i} + \frac{ES_i^2}{2n_i}$$

where $r_i$ is the observed correlation of the scores at times T1 and T2. Note that standardization in the raw score metric yields an effect size that is *not* influenced by the degree of correlation between the scores at T1 and T2 (although the sampling variance of the effect size estimate is).

### *Change Score Metric*

A second option is to standardize the mean change by the standard deviation of the change scores, yielding the effect size

$$\theta_i = \frac{\mu_i^{T2} - \mu_i^{T1}}{\sigma_i^D}$$

where $\sigma^D{}_i$ denotes the standard deviation of the change scores. The corresponding effect size estimate is given by

$$ES_i = \left(1 - \frac{3}{4m_i - 1}\right)\frac{\bar{x}_i^{T2} - \bar{x}_i^{T2=1}}{s_i^D}$$

where $s^D{}_i$ is the observed standard deviation of the change scores (i.e., the standard deviation in the scores after subtracting the T1 score from the T2 score). The sampling variance can be calculated with

$$v_i = \frac{1}{n_i} + \frac{ES_i^2}{2n_i}$$

It can be shown that

$$\theta_i = \frac{\mu_i^{T2} - \mu_i^{T1}}{\sigma_i^D} = \frac{\mu_i^{T2} - \mu_i^{T1}}{\sigma_i\sqrt{2(1 - \rho)}}$$

which reveals that standardization in the change score metric yields an effect size that *is* influenced by the degree of correlation between the scores at T1 and T2. Specifically, when the correlation is greater than .5, then standardization in the change score metric yields a larger effect size than that obtained through standardization in the raw score metric, and vice versa. For more details on the different methods of standardization in the dependent samples case, see Morris and DeShon (2002).

## Correlation Coefficient

The SMD is typically used as the effect size index when interest is centered on the mean difference between two sets of scores (whether from two independent groups or from the same group at two time points). However, in other cases, interest is focused on the strength of the relationship between two continuous variables, in which case the correlation coefficient is usually employed as the effect size measure. Suppose that pairs of scores are obtained within each of the $k$ studies and let $\rho_i$ denote the correlation between the two sets of scores. Now the effect size is defined simply as

$$\theta_i = \rho_i$$

An estimate of $\theta_i$ is given by the raw product–moment correlation coefficient observed in the $i$th study, denoted by $r_i$. It turns out that $r_i$ actually underestimates $\theta_i$ slightly, but this bias can be easily corrected (Olkin & Pratt, 1958) by using

$$ES_i = r_i + \frac{r_i(1 - r_i^2)}{2(n_i - 4)}$$

as the effect size estimate. The sampling variance of $ES_i$ can be computed with

$$v_i = \frac{(1 - ES_i^2)^2}{n_i - 1}$$

The distribution of the raw correlation coefficient becomes increasingly nonnormal as $r_i$ increases. Therefore, several researchers (e.g., Hedges & Olkin, 1985; Lipsey & Wilson, 2001; Rosenthal, 1991) have recommended the use of Fisher's variance stabilizing transformation when meta-analyzing correlation coefficients. Specifically, one computes

$$ES_i = \frac{1}{2}\ln\left[\frac{1 + r_i}{1 - r_i}\right]$$

where ln[ ] denotes the natural logarithm. The sampling variance of $ES_i$ is now given by

$$v_i = \frac{1}{n_i - 3}$$

The advantage of the transformed correlation coefficient is that its distribution is much closer to that of a normal distribution.

Given our experiences in coding studies and attempting to extract effect size information from studies, we cannot help but editorialize a bit at this stage. We would like to appeal to researchers to be more giving of their data and not to forget the archival role of journals. Too many researchers fail to report basic descriptive statistics. These are critical to the meta-analyst and the future researcher interested in comparing your sample to subsequent samples focusing on similar issues. Furthermore, too many researchers report incomplete statistics. For example, one bad habit is to report that the effects were "statistically significant (all $p$'s < .05)." This style of reporting typically leaves the meta-analyst no choice but to throw your article out of the database. Another egregious example is to report findings in graphical form without providing accompanying statistics (means *and* standard deviations). Our least favorite example of this approach is for some authors to report differences as pluses, minuses, or zeros, depending on their idiosyncratic interpretation of the effect size and whether it was positive or negative. Please, please, please report your descriptive statistics and point estimates.

On the surface, coding appears to be a straightforward task. Unfortunately, it is very complicated for some topics and requires extensive coder training. A process we have found useful is to initially have all coders code the same set of five to seven articles and then come together in a meeting to discuss coding discrepancies. The training articles should be selected to have codeable data. This process continues until the structure and content of the coding sheet has stabilized and the number of coding errors reaches an acceptable lower limit. During the process a coding manual is created that specifies how coding decisions are to be made. Random checks of coding then occur after the initial training meetings. These checks can be done by independently coding the article in question and comparing those results with the initial coded results.

To summarize, a coding protocol is a standardized tool that imposes some order on a process that can be rather unruly. The success of a protocol (and a meta-analysis) requires careful planning and an examination of the collected body of studies to determine which information is consistently available for coding. Clear decisions must be made early on to avoid confusion and missed analytic opportunities.

## Preparing the Data for Analysis

Now that you have a database in hand, there are a few additional details to consider. Typically, the data are not in a form that can be readily analyzed, for a variety of reasons that need to be addressed. In particular, the effect sizes most likely need to be converted into a common metric. They may also need to be "directionalized," as, depending on the way a predictor is scored (e.g., as neuroticism or emotional stability), two effects may mean the same thing but have the opposite signs. It is also common to have multiple effect sizes from each sample, and this raises nonindependence issues. Finally, you will need to decide whether to correct for artifacts. We discuss each of these issues in turn.

It is frequently the case that studies report results using a wide range of statistics. Many effect sizes can be converted from one form to another. These include correlation coefficients, standardized mean differences, chi-square statistics obtained from 2 × 2 tables, odds ratios, frequency tables, $t$-tests, $F$-tests, phi coefficients, point-biserial correlations, and means and standard deviations. Moreover, exact $p$-values can be transformed into an effect size if the sample size is known (Rosenthal & Rubin, 2003). There is almost no case in which a bivariate test statistic or outcome measure cannot be transformed into a standard effect size. Your goal at this stage is to simply transform the plethora of effect sizes and test statistics into one common effect size for the analysis stage. To facilitate this process, we have reproduced the formulas for transforming the most common test statistics reported in personality research (see Table 36.2; Rosenthal, 1991).

Certain types of effect sizes may be more difficult to incorporate into a meta-analytic framework. Specifically, partial correlations and beta-weights from complex multiple regression analyses pose a significant challenge. The problem lies in the fact that sampling distributions for each particular type of model would need to be known and converted to a common metric. One solution is to use the $r_{\text{equivalent}}$ statistic in which the $p$-value associated with the test statistic is transformed into a correlation coefficient (Rosenthal & Rubin,

**TABLE 36.2. Common Statistical Transformations in Meta-Analytic Research**

| Statistical score(s) | Transformation |
|---|---|
| | Transformation to Cohen's *d* |
| Means and standard deviations of two groups | $d = \frac{M_1 - M_2}{\sigma_{pooled}}$, $\sigma_{pooled} = \sqrt{\left[\frac{(\sigma_1^2 + \sigma_2^2}{2}\right]}$, or when $\sigma_1 \approx \sigma_2$, then $\sigma_{pooled} \approx \frac{\sigma_1 + \sigma_2}{2}$, i.e., the simple average |
| *t* score with *df* | $d = \frac{2t}{\sqrt{df}}$, when $n_1 = n_2$; or when $n_1 \neq n_2$ $d = \frac{t(n_1 + n_2)}{\left[\sqrt{(df)}\sqrt{(n_1 n_2)}\right]}$ |
| *F* with *df* = 1 in numerator | $r = \sqrt{\frac{F(1,-)}{F(1,-) + df_{error}}}$ |
| *r* | $d = \frac{2r}{\sqrt{1 - r^2}}$ |
| | Transformation to *r* |
| *d* with two known group sizes | $r = \frac{d}{\sqrt{\left[d^2 + \left(\frac{1}{pq}\right)\right]}}$, where $p = \frac{n_1}{N}$ and $q = 1 - p$, or when $p \approx q$, use $r = \frac{d}{\sqrt{d^2 + 4}}$ |
| *p*, converted with *Z*-value table | $r = \frac{Z}{\sqrt{N}}$ |

*Note.* *p* is the proportion of the total sample (*N*) in the first of two groups ($n_1$) being compared.

2003). In some cases, the test statistics from these more complex, multivariate models can be broken into bivariate effect sizes. For example, forward or backward regression analyses often provide sufficient information to permit recovery of the original correlation matrix. A second consideration is using correlations to represent relationships from dichotomous variables as the magnitude of the correlation is sensitive to cell frequencies or base rates. As a result, it is often critical to consider base rates or cell frequencies when converting effect sizes. Another option would be to use odds ratios as a common metric, as they are not biased by base rates or cell frequencies. Two helpful works are Cohen (1988) and Rosenthal (1994).

Another data transformation issue common to personality psychology is the *direction of the effect* problem. For example, one researcher may report the relationship between neuroticism and positive emotionality as –0.50. A second researcher may report the relationship between emotional stability and positive emotionality as 0.50. These correlations, though opposite, reflect the same relationship. Yet, if combined without consideration of the effect direction, the resulting average effect size would be zero and we would erroneously conclude that the domain of neuroticism/emotional stability was unrelated to positive emotionality. The solution to this very common dilemma is to "directionalize" one's effect sizes. In this case, the analyst chooses one particular direction for the relationship and makes sure, by changing the sign where necessary, that the effect size estimates properly represent the appropriate direction. For example, the analyst could choose to represent the relationship as "positive" with "positive" or emotional stability with positive emotionality. This would mean that the effect sizes from any study that reported the relationship between neuroticism and positive emotionality would have to be multiplied by –1 to reverse the sign of the effect size. The meta-analyist will need to be careful about using this method when there is disagreement about the nature of a trait (e.g., positive and negative affect).

In personality psychology one is often confronted with the problem that multiple effect sizes are derived from a single sample. It is quite common for researchers to report the cor-

relation between a simple outcome, such as tobacco consumption, and the entire set of scales drawn from a personality inventory. Even if you are interested in just one domain, such as conscientiousness, most personality inventories contain at least a handful of scales tapping that and other domains. It is problematic to ignore the dependency between these measures drawn from the same sample and can lead to biased estimates of the population parameters.

There are several strategies that can be used to address the dependency between effect sizes. One can randomly select effect sizes from each study, so that any given sample contributes only one effect size to the meta-analysis. Sometimes more systematic selection may be in order. For example, when examining the effect of study moderators on mean-level change in personality traits, we used a strategy in which underrepresented age periods were emphasized rather than randomly selecting from the database (Roberts et al., 2006a). A third solution is to aggregate effect sizes within the sample. We have used this strategy several times to good effect (Bogg & Roberts, 2004; Kuncel, Hezlett, & Ones, 2001; Roberts & DelVecchio, 2000). Although a tremendous number of specific data points are thrown out when using one of these methods, doing so typically yields a more conservative estimate of the population effect sizes. The critical ingredient to a successful aggregation is how studies, samples, and moderator variables are coded, as described above. You will want to anticipate having to rely on one of these strategies by incorporating numbered codes for all of these variables, which can then be used to aggregate the data.

The ideal technique for addressing stochastically dependent effect sizes is to run some form of multivariate analysis in which the correlation among the effects taken from the same sample is accounted for (Gleser & Olkin, 1994). In principle, this is a relatively straightforward procedure. In the example given above in which four conscientiousness measures are used to predict tobacco consumption, all one would need is the correlation among those four conscientiousness measures in that sample. Unfortunately, as we have noted, most researchers fail to include the descriptive statistics relevant to their analyses, and almost no researchers include ancillary analyses, such as the entire correlation matrix of the measures used in the study. One hopes that with the advent of more fluid online publishing of scientific articles and the availability of increased computer storage capacity, supplementary information such as this can be included with research reports as appendixes. Nonetheless, if this information can be acquired, the multivariate approach should be attempted in order to maximize the use of all the information available from each study.

The effect of sampling error on the variability of effect sizes is widely recognized across meta-analytic methods. Less frequently considered is the role of other statistical artifacts. Two loosely defined schools of thought have developed around this issue. The first school is agnostic. The decision of whether to account for artifacts besides sampling error is left up to the researcher (see Cooper & Hedges, 1994). The second school prescribes that as many artifacts as possible should be accounted for because ignoring them may lead to the erroneous conclusion that moderators exist and that the effects vary systematically (Hunter & Schmidt, 2004).

When addressed, the most common of these statistical artifacts are independent variable (IV) and dependent variable (DV) measurement unreliability, IV and DV range restriction, and dichotomization of study variables. All of these artifacts have two important effects on meta-analytic findings. The first is a reduction or attenuation of effects except in the case of range enhancement, which increases the effect. In other words, unreliability, range restriction, and dichotomization of variables reduce the magnitude of observed effects, leading to the conclusion that personality variables have weaker relations with other variables than is actually the case. The second effect is an increase in observed study variability. This has the undesirable effect of potentially leading researchers to believe that results are inconsistent across studies for substantive rather than artifactual reasons. When unaddressed, these artifacts can also make comparisons from study to study, and even meta-analysis to meta-analysis, difficult. A few examples may help to illustrate common situations in which these effects might occur in personality psychology.

The most common scenario is a meta-analysis of the association between a common trait, say, Harm Avoidance, and job performance across a number of different personality measures that all seem to capture the trait. Two problems would arise when some of these measures are markedly less reliable than others. First, the average correlation obtained will tend

to be lower than what would be obtained had all of the studies employed highly reliable measures. Second, the variability across studies will be larger, owing to the measurement properties of the studies rather than substantive moderators.

Another common scenario is a meta-analysis in which the samples differ in their variability on the trait of interest. For example, a meta-analysis of the personality trait of socialization may contain studies done on National Merit Scholar finalists, criminals, and high school students. We would expect the first two samples to be less variable than the third. As in the reliability situation, we will have attenuated effects and increased effect size variability. That is, we might conclude that socialization is not as strongly associated with, say, grades in a course, simply because we have studies that have samples with relatively narrow ranges of socialization.

Artificially dichotomizing a variable, for example, by splitting a sample into a high and low socialization group, results in a loss of information and also tends to reduce the overall observed effect (Cohen, 1990). When samples are dichotomized, all individuals in each half are effectively treated as having equal scores on the dichotomized variable. Again, the meta-analysis would contain estimates of artifactually small effects and increased variability.

The most complete set of methods for addressing these issues was suggested by Hunter and Schmidt (2004) as part of their psychometric meta-analytic approach. This method is especially useful for meta-analytic work when the study samples have been restricted in range owing to direct or indirect selection on the independent variable and/or when the measures across studies vary considerably in reliability. Both of these situations are quite common in personality research, and it is no surprise that the Hunter and Schmidt method has seen extensive use in studies on personality (e.g., Barrick & Mount, 1991; Bono & Judge, 2004; Ones, Viswesvaran, & Schmidt, 1993).

Statistical artifacts can be handled with one of two different approaches. The first is to directly correct each study for its artifacts (i.e., range restriction, unreliability, and dichotomization) and then conduct a meta-analysis with the corrected correlations or standardized mean differences. This approach is ideal in theory but nearly impossible in practice. Rare is the literature in which all studies provide complete information including reliability and variance information. We often feel lucky if a study presents an effect size and information regarding the sample size, let alone local reliability and variance information.

Instead, artifact distributions are commonly used to correct for artifacts. Artifact distributions make use of available information to correct all of the effects in the meta-analysis. The underlying assumption is that the available artifact data represent a reasonable random sample of all the artifact data for all of the studies. This assumption is, of course, more or less tenable across literatures. There are several technical treatments of the artifact distribution method.

In the artifact distribution method, all artifacts and their frequencies are compiled and each study effect is corrected by all possible combinations of the statistical artifacts weighted by their relative frequencies. For example, if the unreliability estimate of 0.70 occurs four times in the database because four studies used the same measure, whereas the unreliability estimate of 0.80 occurs eight times because eight studies used a different measure, the study effects will be corrected by both reliability estimates but the 0.80 corrections will receive twice the weight because they occur twice as often. This method makes maximum use of the available information and allows the researcher to account for the simultaneous effects of both range restriction and unreliability. Unfortunately, this is a more complicated process than simply creating columns of study artifacts. The details become perilous when we need to decide which artifact data to use and how best to use them.[3]

Should you or should you not correct for artifacts? As is true of many methodological choices we make, there are tradeoffs for both approaches. If you choose not to correct for artifacts, it would be prudent to keep in mind that the resulting variability in effect sizes may not be due to moderators. In other words, you should not imbue estimates of heterogeneity with too much significance. However, correcting for artifacts provides somewhat idealized estimates. That is, the corrected estimates reflect what would happen in a world in which we used perfectly reliable measures, optimal sampling techniques, and appropriate measurement models (i.e., not dichotomizing our measures). In this case, one should not imbue the actual magnitude of the population estimates

with too much significance, as researchers in the trenches of flawed primary data collection will most likely never encounter the effect sizes reported in these meta-analyses. Of course, the obvious and common compromise is to report both types of estimates so that the readers can judge for themselves whether an effect exists and what it would look like in an optimal situation.

## How to Analyze Your Data

When conducting a meta-analysis, most of the time will be spent on literature searches, the retrieval of studies, assessing the relevance and quality of the retrieved studies, extraction of effect size estimates, and coding of moderator variables. In this section we discuss how to analyze the data once all of the previous steps have been completed. Although the data analysis takes comparatively little time, some important decisions must be made at this point that can greatly influence the results obtained from the meta-analysis. In particular, there has been an ongoing debate in the literature about the appropriate model to adopt when conducting a meta-analysis, and this is the first issue we address.

### Fixed Effects, Random Effects, Mixed Effects: Which Model Do I Use?

Assume that a collection of $k$ studies has been selected for inclusion in the meta-analysis and that a single (independent) effect size estimate is extracted from each study. Let $ES_i$ denote the $i$th effect size estimate ($i = 1, \ldots, k$). The $ES_i$ values may be, for example, standardized mean differences, raw correlation coefficients, or correlation coefficients after using Fisher's variance stabilizing ($r$ to $z$) transformation. Regardless of the effect size measure used, it is important to recognize that each effect size estimate $ES_i$ is an *estimate* of a corresponding parameter $\theta_i$, which indicates the true effect size in the $i$th study. Therefore, we must draw a clear distinction between the actual or true "effect size" $\theta_i$ and the corresponding "effect size estimate" $ES_i$. Symbolically, this can be expressed by writing

$$ES_i = \theta_i + \varepsilon_i$$

where $e_i$ is the sampling error for the $i$th study. In other words, $ES_i$, the effect size estimate we actually observe in the $i$th study, differs from the true effect size $\theta_i$ by some unknown amount $\varepsilon_i$ simply due to sampling fluctuations. It is usually reasonable to assume that the sampling error $\varepsilon_i$ is normally distributed with mean zero and variance $v_i$.

Not surprisingly, effect size estimates based on larger samples tend to be closer to their corresponding $\theta_i$ values. In other words, effect size estimates based on larger samples have, all else being equal, smaller sampling variances (i.e., smaller $v_i$ values) and therefore should receive proportionally more weight in the analysis because they provide more accurate information. As shown below, we can easily calculate the amount of sampling variance in an effect size estimate. Therefore, corresponding to each $ES_i$ value, we also compute $v_i$, which indicates the amount of sampling variability in the effect size estimate.

Because of the sampling errors, the $ES_i$ values will not coincide across studies. When all of the differences among the effect size estimates can be assumed to be a result of such sampling fluctuations, then the so-called *fixed-effects model* is appropriate. Here, the assumption is that the true effect sizes are exactly the same for all $k$ studies (i.e., $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$), and in this case the effect sizes are said to be *homogeneous*.

However, it is possible (and usually quite likely) that the true effect sizes (i.e., $\theta_1, \theta_2, \ldots, \theta_k$) differ from each other. In that case, the effect sizes are said to be *heterogeneous*. Heterogeneity can be the result of systematic moderator effects, random differences between the true effect sizes, or a combination of both. Depending on the presence of these effects, a more complex model applies.

First, consider the case in which moderators are introducing systematic differences between the effect sizes. For example, in a meta-analysis on social loafing (the tendency of individuals to reduce their effort when working in a group), it was found that the effect size (the difference in performance when effort was evaluated individually versus collectively) depended on the size of the group, with more social loafing occurring as group size increased (Karau & Williams, 1993). Group size, therefore, was a relevant moderator, which differed between the various studies included in the meta-analysis, and therefore should be taken into consider-

ation in the analysis. The appropriate model in this case is the *fixed-effects with moderators model.*

Effect sizes may also differ from each other not because of systematic differences introduced by moderator variables, but owing to random heterogeneity. In this case, the typical assumption is that the $\theta_i$ values are randomly drawn from a normal distribution with mean $\mu_\theta$ and variance $\tau^2$. The size of $\tau^2$ then indicates the amount of random heterogeneity among the effect sizes, and $\mu_\theta$ indicates the average true effect size. The appropriate model in this case is the *random-effects model.*

Finally, it is possible that a combination of systematic moderator effects plus some additional random (residual) heterogeneity are jointly introducing differences into the $\theta_i$ values. In other words, the effect sizes vary systematically with some study-level characteristics and additional heterogeneity exists among the effect sizes that is purely random. The appropriate model in this case is the *mixed-effects model.*

A summary of these four models is given in Table 36.3. To reemphasize the main implications of the various models, imagine that each study included in the meta-analysis on social loafing used a very large sample size (e.g., thousands of subjects). As discussed earlier, the amount of sampling variability in an effect size estimate decreases with the sample size. Consequently, when sample sizes are very large, the sampling variability in each effect size estimate will be so small as to be almost negligible. Therefore, if the fixed-effects model holds, then each $ES_i$ value should be essentially equal to each other and equal to the true population effect size $\theta$. This idea is illustrated in Figure 36.1(a), which shows a plot of 10 hypothetical effect size estimates under the fixed-effects model where $\theta$ is assumed to be .44 (the average effect size found by Karau & Williams, 1993, in their meta-analysis). The figure illustrates how the effect size estimates are clustered around $\theta$, the homogeneous effect size for all 10 studies.

**TABLE 36.3. Four Meta-Analytic Models**

| Model | Moderators present | Random heterogeneity |
|---|---|---|
| Fixed effects | No | No |
| Fixed effects with moderators | Yes | No |
| Random effects | No | Yes |
| Mixed effects | Yes | Yes |

However, when sample sizes are very large and heterogeneity is present, then each effect size estimate will narrow in on its corresponding $\theta_i$ value. In other words, if the $\theta_i$ values are not all equal to each other because they depend on some moderator (such as group size), then the $ES_i$ values should also not be equal to each other, even if the sample size of each study is very large. Figure 36.1(b) shows effect size estimates for 10 hypothetical studies in which the studies are ordered by group size (with study 1 examining the amount of social loafing in small groups and study 10 examining the amount of social loafing in large groups). Because the sample sizes are very large, the sampling variability of the effect size estimates is very small and the pattern created by the moderator variable becomes clearly visible. Calculating a single overall effect size estimate would be meaningless here, because it would reflect neither the amount of social loafing in small groups, nor the amount of social loafing in large groups.

In the random-effects model, variability also will remain in the effect size estimates when sample sizes become very large. However, the variability will not be systematic, as in the fixed-effects with moderators model. Instead, the $\theta_i$ values will simply differ randomly from each other. Consider Figure 36.1(c), which shows a plot of effect size estimates for 10 hypothetical studies with very large sample sizes. The effect sizes were randomly drawn from a normal distribution with $\mu_\theta = .44$ and variance $\tau^2 = .01$. Note that the individual effect size estimates no longer narrow in on a single value, even though the amount of sampling variability is negligible. Instead, the $ES_i$ values narrow in on their corresponding $\theta_i$ values, which in turn fluctuate randomly around $\mu_\theta$.

Finally, large sample sizes will also fail to remove all of the variability from the effect size estimates when the mixed-effects model holds. Consider Figure 36.1(d), which shows a plot of effect size estimates from 10 hypothetical studies under the mixed-effects model, assuming very large sample sizes. Here, the effect sizes depend on a single moderator (group size) plus an additional source of random variability. Therefore, we do recognize the increasing trend in the effect sizes as a function of the moderator, but the effect size estimates still fluctuate

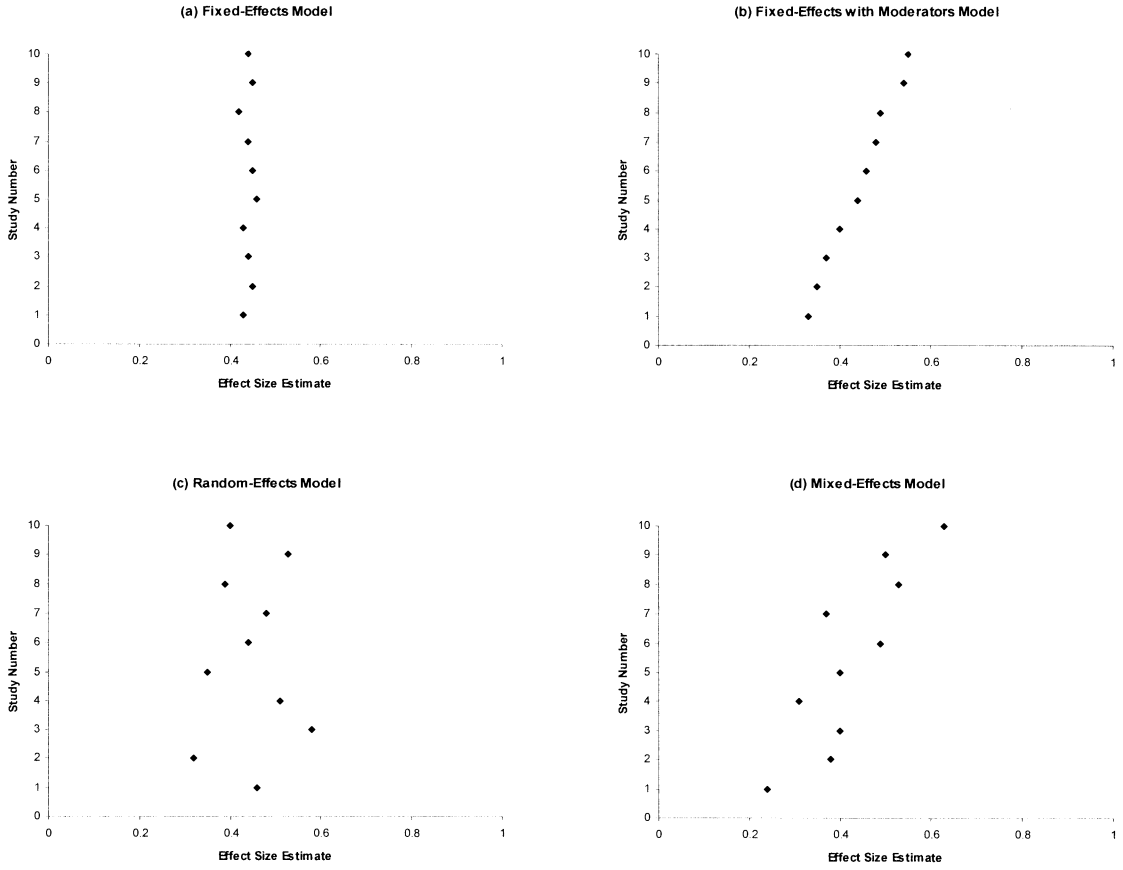


**FIGURE 36.1.** Plot of 10 hypothetical effect-size estimates under fixed-effects model, fixed-effects with moderator model, random-effects model, and mixed-effects model.

noticeably, despite the fact that the amount of sampling variability is negligible.

In practice, the sample sizes of studies included in a meta-analysis are typically not so large that the sampling errors can be disregarded. Patterns like those shown in Figures 36.1(b) and 36.1(d) may then become less discernible to the naked eye. Therefore, it is generally more difficult to determine which of the models is appropriate. We return to this issue below, where we discuss how to approach the model selection task.

## Detecting Publication Bias

Despite his or her best efforts during a literature search, the studies an analyst retrieves from the published literature usually constitutes a subset of all studies that have been conducted on a particular topic. Not surprisingly, studies of higher quality are more likely to be published than those suffering from design flaws or other shortcomings. Although little fault can be found with restricting the published literature to studies of higher quality, the analyst needs to be concerned with the consistent finding that highly statistically significant findings are much more likely to appear in the literature than results that do not reach statistical significance (e.g., Sterling, Rosenbaum, & Weinkam, 1995). For example, researchers may selectively report only those findings that reach significance and/or journal editors/reviewers may favor studies with significance findings. The net effect of this publication bias is that the effect size estimates obtained from the published literature may overestimate the actual effect size. Therefore, publication bias (also called the "file drawer problem") can be a major problem in meta-analysis.

The simplest method for detecting publication bias is by means of a funnel plot. For this,

one plots the effect size estimates against the corresponding sample sizes or variances of the studies. An example of such a plot is shown in Figure 36.2a. Assuming that the fixed-effects model holds, studies with very large sample sizes should fluctuate negligibly around the true θ value. Yet studies with smaller sample sizes should fluctuate more substantially around the true θ value. The figure therefore should look like an inverted funnel (e.g., Figure 36.2b). However, if studies with small effect sizes and small sample sizes (and therefore studies that are unlikely to reach statistical significance) are not published, then the funnel will lack symmetry or will include a hollow area for effect size estimates near zero and small sample sizes. Figure 36.3 illustrates this clearly. This latter shape reflects the fact that as sample size increases the effect sizes move closer to zero, creating the peak of the distribution close to zero. The fact that small studies with small or null effects seldom get published leaves the left portion of the funnel missing. What the researcher must watch out for is the situation in which many small studies with medium effects have been published along with a handful of large studies with very small effects. The small studies with larger effects may lead to mistakenly large population estimates of effect sizes that are due to publication bias, rather than the result of a true effect's occurring.

Visual inspection of funnel plots for publication bias often leaves considerable room for conflicting interpretations, and therefore systematic methods for detecting publication bias have been suggested. Rosenthal (1979), for example, proposed a simple method for calculating the number of unpublished studies averaging null results required to bring the overall level of significance in a research synthesis down to *just significant*. If only a few additional studies with nonsignificant results would be sufficient to do so, then the overall conclusions are argued to be sensitive to publication bias and should be interpreted with caution. However, if hundreds or thousands of studies with null results would be needed, then the findings can be considered robust to publication bias.

More advanced approaches for dealing with publication bias have also been developed. For example, the "trim and fill" method by Duval and Tweedie (2000a, 2000b) allows researchers to estimate the number of studies missing from the published literature due to their not having reached statistical significance and then provides adjusted estimates of the overall effect that account for the missing studies.

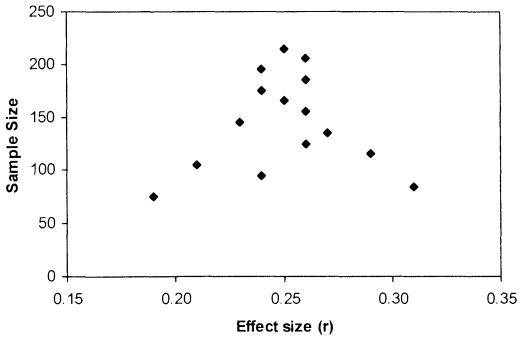


**FIGURE 36.2b.** Example funnel plot for studies with moderate sample sizes (true $r$ = 0.25).

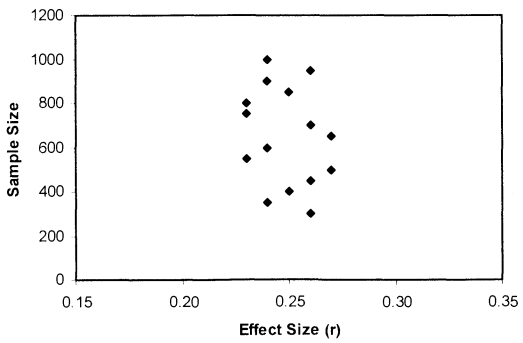


**FIGURE 36.2a.** Example funnel plot for studies with large sample size (true $r$ 0.25).

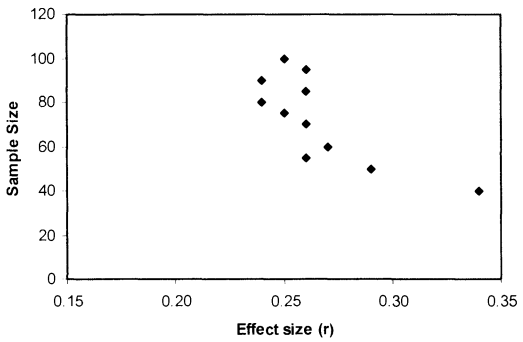


**FIGURE 36.3.** Example funnel plot for studies with small sample sizes (true $r$ = 0.25).

Of course, the most important thing a meta-analyst can do to deal with publication bias is to make a concerted effort to locate the fugitive literature. Statistical adjustments are less than ideal replacements for actual data.

## Testing for Moderators

As indicated earlier, the goal during the data analysis step is to determine which of the four models introduced earlier is most appropriate. Based on recent findings (Viechtbauer, 2004), we advocate starting out with the assumption that the most complex (i.e., the mixed-effects model) is actually the most appropriate model, followed by an examination of possible reductions in model complexity. Therefore, one starts out with the assumption that moderators are present and allows for the possibility that the moderators to be included in the analysis may not account for all of the heterogeneity among the effect sizes.

If no moderators are found to influence the effect sizes, then the question remains whether all of the differences among the effect size estimates are due to sampling fluctuations (in which case we would adopt the fixed-effects model) or whether random heterogeneity is present (in which case, we would adopt the random-effects model). If moderators are found to influence the effect sizes, however, then the question remains whether we can account for all of the differences among the effect sizes based on the moderators (in which case we would adopt the fixed-effects with moderators model) or whether there is residual heterogeneity in the effect sizes (in which case, we would adopt the mixed-effects model).

Fitting the mixed-effects model requires specifying a model for the relationship between the effect sizes and the moderators, estimating the parameters of this model, and estimating the amount of residual heterogeneity in the effect sizes (i.e., the amount of heterogeneity in the effect sizes that is not accounted for by the moderators). Methods for fitting the mixed-effects model have been described in the literature (e.g., Konstantopoulos & Hedges, 2004; Overton, 1998; Raudenbush, 1994; Raudenbush & Bryk, 1985, 2002; Sheu & Suzuki, 2001; Viechtbauer, 2004) and are beyond the scope of this chapter. Although working with mixed-effects models requires more statistical expertise on the part of the meta-analyst than using the still popular fixed-effects models, we want to stress here that moderator tests should be conducted in the context of a mixed-effects model. It can be shown (Viechtbauer, 2004) that the Type I error of moderator tests in the context of fixed-effects models can become severely inflated, leading researchers to discover spurious moderators (i.e., moderators that are actually unrelated to the effect sizes often turn out to be significant when tested with fixed-effects models). However, the mixed-effects model adequately controls the Type I error rate and therefore should be preferred.

If none of the moderators turn out to be significant, then a single overall effect size can be provided. A random-effects model should be preferred in this case over the fixed-effects model, unless the evidence suggests (usually by means of a homogeneity test) that the effect sizes are truly homogeneous. For more details, see Shadish and Haddock (1994).

## Aggregating Effect Sizes

If the effect sizes are not influenced by moderators, then a single aggregated effect size estimate provides an adequate summary of the data.[4] In that case, we must still distinguish between two cases, namely, whether the effect sizes are heterogeneous because of random differences (i.e., the effect sizes are assumed to be randomly drawn from a normal distribution with mean $\mu_\theta$ and variance $\tau^2$) or homogeneous (i.e., all the effect sizes are equal to each other: $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$). The random-effects model is appropriate in the former case and the aggregated effect size estimates $\mu_\theta$. However, the fixed-effects model applies to the latter case and we estimate $\theta$.

If the effect sizes are homogeneous, then an estimate of $\theta$ is given by

$$\hat{\theta} = \frac{\sum w_i ES_i}{\sum w_i}$$

where $w_i = 1/v_i$. Therefore, studies with larger sample sizes (and consequently smaller sampling variances) are given more weight, as they provide more accurate information about the true effect size $\theta$. One can also obtain an approximate 95% confidence interval for $\theta$ with

$$\hat{\theta} \pm 1.96 \sqrt{\frac{1}{\sum w_i}}$$

as a way to gauge the precision of the estimate of $\theta$. If the confidence interval includes zero,

then this is equivalent to testing $H_0$: $\theta = 0$ at $\alpha = .05$ and failing to reject the null hypothesis that the effect size is equal to zero.

One can test whether the effect sizes are actually homogeneous (i.e., whether the fixed-effects model is appropriate for the data) with the so-called $Q$-test by computing

$$Q = \sum w_i (ES_i - \hat{\theta})^2$$

If $Q$ exceeds the critical value of a chi-square distribution with $k - 1$ degrees of freedom, then this suggests the presence of heterogeneity among the effect sizes. An estimate of the amount of heterogeneity among the effect sizes (DerSimonian & Laird, 1986) is then given by

$$\hat{\tau}^2 = \frac{Q - (k-1)}{c}$$

where

$$c = \sum w_i - \frac{\sum w_i^2}{\sum w_i}$$

An estimate of $\mu_\theta$ is then obtained with

$$\hat{\mu}_\theta = \frac{\sum w_i ES_i}{\sum w_i}$$

where $w_i = 1/(v_i + \hat{\tau}^2)$. An approximate 95% confidence interval for $\mu_\theta$ is given by

$$\hat{\mu}_\theta \pm 1.96 \sqrt{\frac{1}{\sum w_i}}$$

again with $w_i = 1/(v_i + \hat{\tau}^2)$. This confidence interval will always be wider than the one computed under the fixed-effects model, reflecting the additional variability introduced by the heterogeneity among the effect sizes. Inclusion of zero in the interval indicates that we cannot reject $H_0$: $\mu_\theta = 0$ at $\alpha = 0.05$.

Because the $Q$-test is not infallible, it is generally advisable to automatically adopt the random-effects model and to estimate the amount of heterogeneity as described above. Should the effect sizes be homogeneous, then $\hat{\tau}^2$ will tend to be close to zero or even negative, in which case $\hat{\tau}^2$ is truncated to zero and the random-effects model reduces to the fixed-effects model (note that the equations for estimating $\theta$ and $\mu_\theta$ differ only in the weights used, which are identical when $\hat{\tau}^2 = 0$). A pragmatic approach would be to compute the aggregated effect size and corresponding confidence interval under both the fixed- and the random-effects model, as part of a sensitivity analysis.

In sum, analysis of the meta-analytic database proceeds very quickly once the appropriate model is chosen. The key choices are whether one believes that the population effect sizes are heterogeneous and how to account for moderators. Once the analyses have been computed, then the appropriate findings should be reported. Common approaches to this step of the process are described in Halvorsen (1994).

## Lessons Learned

Although we are by no means the most prolific users of meta-analytic techniques, we have done enough of them now to provide some insights and lessons that may help the budding personality meta-analyst.

First, the most important lesson we have learned is that meta-analyses are a lot of work. The ignorant sap that maligns your meta-analytic efforts as easy because they entail secondary analyses of already collected data should be scolded. To do an exhaustive and therefore authoritative meta-analytic review will typically take a few years, minimum. Keep this in mind when planning your meta-analysis.

A second lesson we have learned concerns the blessings and banes of the Big Five taxonomy of personality traits. In some ways the Big Five are a godsend to meta-analysts. The ability to organize the dizzying array of personality measures post hoc into five basic domains has allowed for key findings in many areas of psychology and related fields. Without the Big Five we would not understand which personality traits are most important for specific types of job outcomes (Hogan & Holland, 2003), creativity (Feist, 1998), and criminal behavior (Miller & Lynam, 2001). The pre-Big Five meta-analyst was forced to provide an estimate for "personality," rather than for different traits within the personality trait taxonomy, which essentially washed out differential relationships. In this respect, the Big Five is a wonderful tool and has provided invaluable clarity on the role of personality traits in numerous domains.

We are also acutely aware of the limitation of the global categorization of traits into these five domains. In many regards, the Big Five are too broad, and very specific facets of each domain are more interesting and theoretically rel-

evant. In our study tying conscientiousness to health behaviors, for instance, we found certain facets to be much more important than others in predicting health behaviors. For example, the traditionalism and impulse control facets had much more pervasive effects across health domains, whereas the organization facet did not. We demonstrated similar refinements in our analysis of mean-level change in personality. Based on the work of Helson and Kwan (2000), we reorganized the domain of extraversion into the subdomains of social dominance and social vitality. This led to starkly different findings. People increased substantially on measures of social dominance and showed little or no change on measures of social vitality. If we had simply merged the two facets into one overall domain of extraversion, these patterns would have gone undetected. To refine our categorization of traits, we need a taxonomy of personality traits that is more specific than the Big Five (see Roberts, Walton, & Viechtbauer, 2006b). Unfortunately, there exists no empirically supported lower-order taxonomy of personality traits. We hope that this taxonomy is identified in the near future so we can further refine our measurement of personality and our organization of meta-analyses of personality traits.

One of the key lessons we have learned from doing meta-analysis is the insanity of null hypothesis significance testing. Doing a meta-analysis makes one acutely aware of the importance of effect sizes and the capricious nature of statistical significance as an arbiter of whether someone claims an effect is present or not. By looking across studies that vary in terms of their sample size, you are automatically confronted with the fact that the studies with 50–100 people have to contain medium to large effects in order to satisfy the typical null hypothesis significance testing (NHST) standards. In turn, these same studies essentially throw away effect sizes below 0.20. At the same time, researchers who bother to collect a respectable number of data points lay claim to effects as statistically significant that are below 0.20. The beauty of meta-analytic techniques is that by combining data across multiple studies, statistical significance becomes a moot point. Everything is statistically significant from zero, making statistical significance an uninteresting and uninformative standard by which to judge whether an effect is real.

In terms of the typical approach to designing our studies, we would like to get in line behind the many researchers who have noted that our studies lack the power to detect the effects we are interested in. Too often, we use rules of thumb to determine how many participants to include in our research without planning or thinking about the number that would be appropriate, given the magnitude of the effects we expect to find. We have a sweeping recommendation to make. Regardless of the number of participants you are planning to incorporate in your study, double it. That way, like meta-analysis, poor power will not deter the science of personality from accumulating meaningful patterns of results.

In conclusion, we hope that our overview of meta-analysis in personality psychology is both helpful and informative. As we noted at the beginning of this chapter, meta-analyses can bring clarity to research domains that often appear at first blush to be muddled and confused. They also tend to shift the question ever so slightly from "Is there an effect?" to "What size is the effect?"—a shift that we believe can better lead to a science of personality psychology that stands much more firmly in the face of criticism and that cumulates findings in a more productive fashion.

## Acknowledgment

Preparation of this chapter was supported by Grant No. R01 AG21178 from the National Institute on Aging.

## Notes

1. This task is preferably carried out by a highly motivated researcher/graduate student or an assistant with obsessive–compulsive tendencies.
2. We have found one special circumstance when null findings have been preferred. Specifically, researchers have been prone to overreport null findings for the validity of achievement tests such as the SAT (Hezlet et al., 2001), whereas positive results tend not to be published.
3. For an in-depth treatment of these issues, see Hunter and Schmidt (2004).
4. It is important to realize that a single aggregated effect size estimate can be misleading when moderators are actually present. Take, for example, the situation in which the effect sizes depend on a single dichotomous moderator and the effect sizes

are negative (e.g., –0.5) for half of the set of studies and positive (e.g., +0.5) for the other half. An aggregated effect size estimate would then be close to zero, suggesting the absence of an effect.

## Recommended Readings


Cooper, H. M., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Sage.

Duval, S. J., & Tweedie, R. L. (2000a). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.


## References


Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin, 130*, 887–919.

Bono, J. E., & Judge, T. (2004). Personality and transformational and transactional leadership: A meta-analysis. *Journal of Applied Psychology, 89*, 901–910.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312.

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

Cooper, H. M., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Sage.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177–188.

Duval, S. J., & Tweedie, R. L. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–98.

Duval, S. J., & Tweedie, R. L. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463.

Feist, G. J. (1998). A meta-analysis of personality in scientific and artistic creativity. *Personality and Social Psychology Review, 2*, 290–309.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Sage.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–356). New York: Sage.

Halvorson, K. T. (1924). The reporting format. In H. M. Cooper & L. V. Hodges (Eds.), *The handbook of research synthesis* (pp. 425–438). New York: Russell Sage.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Helson, R., & Kwan, V. S. Y. (2000). Personality development in adulthood: The broad picture and processes in one longitudinal sample. In S. Hampson (Ed.), *Advances in personality psychology* (Vol. 1, pp. 77–106). London: Routledge.

Hezlett, S. A., Kuncel, N. R., Vey, M. A., Ahart, A., Ones, D. S., Campbell, J. P., et al. (2001, April). The predictive validity of the SAT: A comprehensive meta-analysis. In D. S. Ones & S. A. Hezlett (Chairs), *Predicting Performance: The Interface of I/O Psychology and Educational Research*. Symposia presented at the annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*, 100–112.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Newbury Park, CA: Sage.

Karau, S. J., & Williams, K.D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*, 681–706.

Konstantopoulos, S., & Hedges, L. V. (2004). Meta-analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 281–297). Thousand Oaks, CA: Sage.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The reliability of self-reported grade point averages, class ranks, and test scores. *Review of Educational Research, 75*, 63–87.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162–181.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment. *American Psychologist, 56*, 128–165.

Miller, J. D., & Lynam, D. (2001). Structural models

of personality and their relation to antisocial behavior: A meta-analytic review. *Criminology, 39*, 765–798.

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.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29*, 201–211.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679–703.

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

Raudenbush, S. W. (1994). Random effects models. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Sage.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*, 75–98.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126*, 3–25.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006a). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132*, 1–25.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006b). Personality traits change in adulthood: Reply to Costa & McCrae (2006). *Psychological Bulletin, 132*, 29–32.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Sage.

Rosenthal, R., & Rubin, D. B. (2003). R-equivalent: A simple effect size indicator. *Psychological Methods, 8*, 492–496.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Sage.

Sheu, C.-F., & Suzuki, S. (2001). Meta-analysis using linear mixed models. *Behavior Research Methods, Instruments, and Computers, 33*, 102–107.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 49*, 108–112.

Viechtbauer, W. (2004). *Model selection strategies in meta-analysis: Choosing between the fixed-, random-, and mixed-effects model*. Manuscript submitted for publication.