9 General Technical Issues in Meta-Analysis

This chapter discusses technical issues that are general to meta-analysis. That is, these issues apply whether meta-analysis is applied to correlations, to *d* values, or to other statistics (e.g., odds ratios). Also, these issues apply whether the methods used are those presented in this book, those of Hedges and Olkin (1985), Borenstein et al. (2009), those of Rosenthal (1984, 1991), or any other methods. The issue of fixed versus random effects meta-analysis models is general in nature but is not included in this chapter because it has been fully addressed in Chapters 5 and 8. New developments in meta-analysis methods occur with some frequency (Schmidt, 1988). Some of these developments are explored in this chapter.

First, we discuss the contention that large-sample studies are a substitute for meta-analysis and show why this view is incorrect. Second, we discuss the various methodological issues involved in detecting moderators (interactions) in meta-analysis, including subgrouping of studies and meta-regression. Next, we introduce second-order sampling error (the sampling error remaining in the results of a meta-analysis), and we present methods for second-order meta-analysis (meta-analysis of meta-analyses) that address some of the problems created by second-order sampling error. We then provide a complete technical treatment of second-order sampling error and its effect on confidence intervals in meta-analysis. In this connection, we point out differences in the way confidence intervals for random effects meta-analyses are computed in the Hedges-Olkin and Hunter-Schmidt methods. Next, the technical issue of how to update a meta-analysis when new studies become available and the question of optimal study weights in meta-analysis are discussed. This is followed by a discussion of a more informative way to view and interpret percent variance accounted for in a meta-analysis. Finally, we present a discussion of a statistical index of effect sizes not treated elsewhere in this book: the odds ratio. Last, we present the reader with three exercises: conducting second-order meta-analysis in two different ways.

Large-*N* Studies Versus Meta-Analysis

Some have argued that the need for meta-analysis is merely a consequence of small-sample studies with their typically low levels of statistical power. The argument is made that researchers should conduct only large-sample studies (i.e., studies with *N*s of 2,000 or more) and that such studies, with their higher statistical power, would make meta-analysis unnecessary (see, e.g., Bobko & Stone-Romero, 1998; Murphy, 1997). We question this position for three reasons: (1) It leads to a reduction in the total amount of information available in the literature for the calibration of correlations and effect sizes, (2) it reduces the ability to detect the presence of potential moderators, and (3) it does not eliminate the need for meta-analysis.

Loss of Information. For practical reasons, many researchers cannot obtain large sample sizes, despite their best efforts. If a requirement for large *N*s is imposed, many studies that would otherwise be conducted and published will not be conducted—studies that could contribute useful information to subsequent meta-analyses (Schmidt, 1996). This is what has happened in the area of validity studies in personnel psychology. After publication of the study by Schmidt et al. (1976) showing that statistical power in traditional validity studies averaged only about .50, the average sample sizes of published studies increased from around 70 to more than 300. However, the *number* of studies declined dramatically, with the result that the total amount of information created per year or per decade (expressed as *N*s in a meta-analysis) for entry into validity generalization studies decreased. That is, the total amount of information generated in the earlier period from a large number of small-sample studies was greater than that generated in the later period for a much smaller number of larger-sample studies. Hence, there was a net loss in ability to calibrate validities.

Reduced Ability to Detect Potential Moderators. The situation described previously creates a net loss of information *even if there are no moderator variables* to be detected, that is, even if $SD_\rho = 0$ in all validity domains studied. Although $SD_\rho = 0$ is a viable hypothesis in the predictor domains of ability and aptitude tests (Schmidt et al., 1993), this hypothesis may not be viable in some other predictor domains (e.g., assessment centers, college grades). And it is certainly not viable in many research areas outside personnel selection. If $SD_\rho = 0$, the total number of studies does not matter; all that matters in determining the accuracy of the meta-analysis study is the total *N* across all studies in the meta-analysis. As described previously, this total *N* has been reduced in recent years. If $SD_\rho > 0$, however, it is critical to have an accurate estimate of SD_ρ. In estimating SD_ρ, *N* is *the number of studies.* Hence, holding the total *N* in the meta-analysis constant, a small number of large studies provides a less accurate estimate of SD_ρ than does a large number of small studies. A large number of small

studies samples a much more numerous array of potential moderators—in fact, each small study samples different potential moderators that might contribute to $SD_\rho > 0$. For example, suppose total N for the meta-analysis is 5,000. If this total N consists of four studies each with $N = 1{,}250$, then the estimate of SD_ρ is based on *only four data points:* four samples from the distribution of ρ. On the other hand, if this total N consists of 50 studies of $N = 100$ each, then the estimate of SD_ρ is based on 50 data points sampled from the ρ distribution—and is therefore likely to be much more accurate. This greatly increases what Cook and Campbell (1976, 1979) called "external validity."

Bobko and Stone-Romero (1998) argued that this same level of precision of estimation for SD_ρ can be obtained with a single large-N study by, in effect, dividing the one large study into many smaller ones. This is unlikely to be true. The single large study reflects the way a single researcher or set of researchers conducted that one study: same measures, same population, same analysis procedures, and so forth. It is unlikely to contain within itself the kinds of variations in the methods and potential moderator variables that are found in 50 independently conducted studies. Another way to see this is to consider the continuum of different types of replications of studies (Aronson, Ellsworth, Carlsmith, & Gonzales, 1990). In a literal replication, the same researcher conducts the new study in exactly the same way as in the original study. In an operational replication, a different researcher attempts to duplicate the original study. In systematic replication, a second researcher conducts a study in which many features of the original study are maintained but some aspects (e.g., types of subjects) are changed. Literal and operational replications contribute in only a limited way to external validity (generalizability) of findings, but systematic replications are useful in assessing the generalizability of findings across different types of subjects, measures, and so on. Finally, in the case of constructive replications, the researcher attempts to vary most of the aspects of the initial study's methods, including subject type, measures, and manipulations. Successful constructive replication adds greatly to the external validity of a finding. Breaking up a large study into "pieces" is similar to the creation of several smaller literal replications and does not contribute to external validity or generalizability of findings. However, in a meta-analysis of a large number of small studies, the studies in the meta-analysis constitute systematic or constructive replications of each other; that is, many study aspects vary across studies. In these circumstances, a finding of a small SD_ρ (or a small SD_δ) provides strong support for generalizability—that is, this result is strong evidence of external validity of the finding. As discussed in Chapter 4, this finding is common in the personnel selection area. If the number of studies in the meta-analysis is small, even if each study is a large-sample study, the meta-analysis is weaker because the number of systematic or constructive replications underlying the final results is smaller, and hence, external validity is more

questionable. This is another approach to understanding why a large number of small studies are better than a small number of large studies.

Meta-Analysis Still Necessary. Finally, even if all studies conducted are large-sample studies, it is still necessary to integrate findings across studies to ascertain the meaning of the set of studies as a whole. Because meta-analysis is the statistically optimal method for doing this, meta-analysis is still necessary. In concluding that meta-analysis would no longer be necessary, advocates of the position we are critiquing appear to be thinking of the fact that large-*N* studies, with their high statistical power, will show agreement on statistical significance tests: If there is an effect, all studies should detect it as statistically significant. However, this does not mean meta-analysis is unnecessary. What is important is the estimates of effect size magnitudes. Effect size estimates will still vary across studies, and meta-analysis is still necessary to integrate these findings across studies. Hence, we cannot escape the need for meta-analysis.

We conclude therefore that a movement to a smaller number of larger *N* studies would not contribute to the advancement of cumulative knowledge in any area of research. In fact, it would be detrimental to knowledge generation and discovery. And it would not eliminate the need for meta-analysis.

Detecting Moderator Variables in Meta-Analysis

A variety of issues arise when meta-analysis is used to detect moderators (or interactions). One of these issues is capitalization on sampling error when focusing on only those potential moderators that show statistical significance when a larger number of potential moderators are examined. This issue was explored in some detail near the end of Chapter 2. The related issues discussed in this chapter include (a) detecting moderators not hypothesized a priori, (b) use of hierarchical meta-analysis in moderator detection, (c) meta-regression in moderator detection (including "mixed models" of meta-analysis), and (d) multilevel meta-analysis and hierarchical linear models (HLM).

DETECTING MODERATORS NOT HYPOTHESIZED A PRIORI

When the moderator variable is not specified or hypothesized in advance by theory, the statistical power of a meta-analysis with respect to the variance of ρ or δ is the probability that the meta-analysis will detect variation in ρ or δ values across studies when such variation does, in fact, exist. One minus this probability is the probability of a Type II error: concluding that all the variance across studies is due to artifacts when, in fact, some of it is

real. When all variance is indeed artifactually caused, there is no possibility of a Type II error, and there can be no statistical power question. Just as second-order sampling error becomes more of a problem as the number of studies becomes smaller, statistical power also becomes lower. A number of statistical tools have been used to make the decision about whether any of the observed variance is real. In our meta-analytic research on test validities, we have used the 75% rule of thumb: If 75% or more of the variance is due to artifacts, we conclude that all of it is, on grounds that the remaining 25% is likely to be due to artifacts for which no correction has been made. Another method is the chi-square test of homogeneity. As pointed out in Chapters 5 and 8 and again in this chapter, this test has low power under most realistic circumstances (Hedges & Pigott, 2001; National Research Council, 1992). In addition, it has all the other disadvantages of significance tests, as discussed in Chapter 2. Callender and Osburn (1981) presented a third method, one based on simulation.

Extensive computer simulation studies have been conducted to estimate the statistical power of meta-analyses to detect variation in ρ using these decision rules (Aguinis et al., 2008; Osburn, Callender, Greener, & Ashworth, 1983; Sackett, Harris, & Orr, 1986; Spector & Levine, 1987). These estimates have been obtained for different combinations of (1) numbers of studies, (2) sample size of studies, (3) amount of variation in ρ, (4) mean ρ values, and (5) levels of measurement error. The findings of the Sackett et al. (1986) study are consistent with the others and are probably the most relevant to meta-analysis in general. Sackett et al. found that, under all conditions, the 75% rule had "statistical power" greater than (or equal to) the other methods, including the *Q* statistic (although the 75% rule also showed a higher Type I error rate: concluding there was a moderator when there was not). The term *statistical power* is placed in quotation marks here because that term applies only to significance tests, and the 75% rule is not a significance test but rather a simple "rule of thumb" decision rule. The advantage in statistical power for the 75% rule was relatively the greatest when the number of studies was small (4, 8, 16, 32, or 64) and the sample size of each study was small (50 or 100). However, when the assumed population variance to be detected (s^2_ρ) was small, and both the number of studies and the sample size of the studies were small, all methods had relatively low statistical power. For example, if there were four studies ($N = 50$ each) with $\rho = .25$ and four studies ($N = 50$ each also) with $\rho = .35$ (corresponding to $s^2_\rho = .01$), and if $r_{xx} = r_{yy} = .80$ in all studies, statistical power was .34 for the 75% rule and only .08 for the other methods. However, a total sample size of 8(50) = 400 is very small, and 8 is a small number of studies for a meta-analysis. Also, a difference of .10 is very small. If the difference in this example is raised to .30, power rises to .75. This difference between ρs is more representative of the moderators that it would be theoretically and practically important to study. Nevertheless, it is true that individual meta-analyses have less than optimal statistical power in some cases. As the reader of this book is by now aware, we recommend against the

use of significance tests (see, e.g., Chapter 2). These simulation studies show that our simple 75% rule typically is more accurate than significance tests used to assess homogeneity. However, no decision rule for judging homogeneity versus heterogeneity in meta-analyses of realistic sets of studies has perfect accuracy.

The preceding discussion applies to "omnibus" tests for moderator variables—moderator variables that are not specified in advance by theories or hypotheses. In such cases, the existence of moderators must be detected by determining whether the variance of study effect sizes is larger than can be accounted for by the presence of variance-generating artifacts. The story is very different when the moderator hypotheses are specified in advance. In such cases, the studies in the meta-analysis can be subgrouped based on the moderator hypothesis (e.g., studies done on blue- vs. white-collar employees), and credibility and confidence intervals can be placed around the means ($\bar{\delta}$ or $\bar{\rho}$) of the subgroup meta-analyses, as described in earlier chapters (Chapters 3, 4, 5, 7, and 8) and later in this chapter. Confidence intervals are most relevant in assessing moderator variables if the main focus of interest is on mean differences. Credibility intervals are most relevant if the focus of interest is on whole distributions of parameters. This procedure is much more effective in identifying moderators than operating without a priori moderator hypotheses and attempting to assess the presence of moderators by testing for heterogeneity in observed *d* or *r* values.

In most areas of research, there should be sufficient development of theory to generate hypotheses about moderators. However, in one major meta-analytic research area—the generalizability of employment test validities—this has not been the case. It has not been possible to use the subgrouping approach to test the "situational specificity" hypothesis in personnel selection. To use this approach, the moderators must be specified. There must be a theory, or at least a hypothesis, that is specific enough to postulate that, for example, correlations will be larger for females than for males, or larger for "high-growth-need" individuals than for "low-growth-need" individuals, or larger in situations where supervisors are high in "consideration" than where supervisors are low in consideration. The situational specificity hypothesis does not meet this criterion; it postulates merely that there are unspecified subtle but important differences from job to job and setting to setting in what constitutes job performance, and that job analysts and other human observers are not proficient enough as information processors to detect these critical elusive differences (Albright, Glennon, & Smith, 1963, p. 18; Lawshe, 1948, p. 13). When the operative moderators are actually unknown and unidentifiable, it is not possible to subgroup studies by hypothesized moderators. However, if one can show that all observed validity variance is due to artifacts, one has shown that no moderators can possibly be operating. This approach does not require that the postulated moderators be identified or

even identifiable. Given that there is a broad and heterogeneous range of situations represented in one's meta-analysis, one can show that the postulated moderators do not exist, even without knowing what the moderators might be.

However, some might make statements like the following: "There are many factors that could affect outcomes. Supervisory style may have important effects; group membership, geographical location, type of industry, and many other variables would be expected to be moderators." Such statements are usually not based on theoretical reasoning or empirical evidence. They are usually just unsupported speculations and, thus, are not scientifically useful. Because the number of hypothesized potential moderators is essentially unlimited, it will never be possible to test them all using the second, more effective, procedure. However, the first procedure—the omnibus procedure we have used to test the situational specificity hypothesis—can be used to test all such moderators simultaneously, even those that have not yet been named by the critic. If the meta-analysis is based on a *large* group of studies that is heterogeneous across all potential moderators, then a finding that artifacts account for all between-study variance in correlations or effect sizes indicates that none of the postulated moderators are, in fact, moderators. Even when all the variance is not accounted for by artifacts, the remaining variance may often be small, demonstrating that even if some moderators might exist, their effect is far more limited in scope than implied by the critic. In fact, the results may often indicate that the moderators have at best only trivial effects (Schmidt et al., 1993). In this connection, it should always be remembered that the variance remaining after correction for artifacts indicates the *upper bound* of the effects of the moderators. This will almost always be true because, as described in Chapters 3, 4, 5, and 7 there will almost always be some artifacts operating to create variance for which no corrections will be possible.

The facts of second-order sampling error and less than perfect "statistical power" in individual meta-analyses point to another reason for the importance of a principle we stated in Chapters 1 and 2. The results of a meta-analysis should not be interpreted in isolation but rather in relation to a broader set of linked findings from other meta-analyses that form the foundation for theoretical explanations. Estimating a particular relationship is only the immediate objective of a meta-analysis; the ultimate objective is to contribute pieces of information that can be fitted into a wider developing mosaic of theory and understanding. However, just as the results of a meta-analysis can contribute to this bigger picture understanding, so also can the resulting bigger picture understanding contribute to the interpretation of particular meta-analysis results. Results of "small" meta-analyses (those based on few studies and small-sample studies) that are inconsistent with the broader cumulative picture of knowledge thereby become suspect, while the credibility of those that are consistent is enhanced. This is the universal pattern in science of reciprocal causation between data and theory.

Some have worried that the inadequate ability of meta-analysis to detect moderators might be an almost insurmountable problem limiting scientific progress (even while admitting that better alternatives to meta-analysis do not exist). The critical difficulty with this argument is that it focuses on single meta-analytic studies. Just as earlier researchers focused on the individual study, failing to realize that single studies cannot be interpreted in isolation, this position focuses on single meta-analyses—in particular, on the sometimes weak ability of single meta-analyses to identify moderators—not seeing that it is the overall pattern of findings from many meta-analyses that is important in revealing the underlying reality.

Consider an example in which the overall pattern of findings was critical. In personnel selection, the theory of situational specificity holds that the true (population) validity of any employment test varies substantially from one organization to another even for highly similar or identical jobs. This is the hypothesis that $S_\rho^2 > 0$. In meta-analysis (called validity generalization when used in personnel selection), this hypothesis is tested by determining whether artifacts such as sampling error account for the variation of observed validity coefficients across studies conducted in different organizations on similar jobs using measures of the same ability (e.g., arithmetic reasoning). In the initial validity generalization studies, the average percentage of the observed validity variance accounted for by artifacts was less than 100%. However, these meta-analyses were based on published and unpublished studies from a wide variety of sources and researchers, and we pointed out in all our studies that there were several sources of between-study variance that we could neither control for nor correct for (e.g., programmer errors, transcriptional errors; see Chapter 5 and Schmidt et al., 1993). When all studies going into a validity generalization analysis are conducted by the same research team, strong efforts can be made to control these sources of errors. In two large-scale, nationwide consortium studies, such efforts were made (Dunnette et al., 1982; Peterson, 1982). In both cases, these studies found that, on average, all variance across settings (i.e., companies) was accounted for by artifacts. The same was found to be true in data from studies conducted in 16 companies by Psychological Services, Inc. (Dye, 1982). Thus, our prediction that improved control of sources of error variance would show that all between-study variance is due to artifacts was borne out. These findings are strong evidence that there is no situational specificity in the validity of employment tests of cognitive ability.

There were more aspects to the pattern of evidence against situational specificity, however. The situational specificity hypothesis predicts that if the situation is held constant and the tests, criteria, and job remain unchanged, validity findings should be constant across different studies conducted in that setting. That is, because the setting is constant, observed validities should be constant because it is differences between settings that are hypothesized to cause differences in observed validities. Meta-analytic

principles predict that such observed validities will vary substantially, mostly because of sampling error. We tested these predictions in two studies (Schmidt & Hunter, 1984; Schmidt, Ocasio, et al., 1985) and found that observed validities within the same situation varied markedly, disconfirming the situational specificity hypothesis. In the second of these studies, the data from a large-sample validity study (N = 1,455) were divided into smaller, randomly identical studies (21 studies of n = 68 each). Because situational variables were held constant, the specificity hypothesis predicted that all the smaller studies would show the same observed validity. This was not the case, however. Instead, there was great variance among studies in both magnitude of validity and statistical significance level, as predicted by the theory of artifacts, which is the basis of meta-analysis and validity generalization. A key finding was that the variation in validities was as great as that typically found across similar studies conducted in entirely different settings.

The final piece of evidence that fits into this framework is this: Recent refinements in validity generalization methods have led to the conclusion that published validity generalization studies substantially underestimate the percentage of observed validity variance that is due to artifacts, further undercutting the situational specificity hypothesis. There are three such refinements. First, non-Pearson validity coefficients are removed, because the sampling error formula for Pearson correlations substantially underestimates the sampling error in non-Pearson correlations such as the biserial and the tetrachoric (see Chapter 5). Second, within each meta-analysis, the population observed correlation used in the sampling error formula is estimated by the mean observed validity instead of the individual observed validity from the study at hand. This provides a more accurate estimate of sampling error (see Chapter 5). Third, the problem created by nonlinearity in the range restriction correction (cf. Chapter 5 and Law et al., 1994a, 1994b) is solved by a new set of computational procedures. Schmidt et al. (1993) applied these improvements in the massive validity database in Pearlman et al. (1980), which consisted of approximately 3,600 validity coefficients from published and unpublished studies from many organizations, researchers, and more than periods ranging over 70 years. Each of these methodological refinements resulted in increases in the percentage of validity variance accounted for and smaller estimates of SD_ρ. Even in this heterogeneous group of studies, almost all validity variance (nearly 90%) was found to be due to artifacts. This research is discussed in more detail in Chapter 5.

All these pieces of interlocking evidence point in the same direction: toward the conclusion that, for employment tests of cognitive abilities, the situational specificity hypothesis is false. The only conclusion consistent with the total pattern of evidence is that there is no situational specificity (or that situational effects are so tiny that it is reasonable to consider them to be 0; some prefer this latter conclusion, which we regard as scientifically identical).

In some research areas, there may be no related meta-analyses with which one's meta-analytic results can be cross-referenced and checked for consistency. In such cases, one's results should be compared with the broader pattern of general research findings. Where even this is not possible, meta-analyses based on small numbers of studies should indeed be interpreted with caution, even though the meta-analysis provides the most accurate summary possible of existing research knowledge at that point in time. We stress that, in cases such as this, the problem is created not by meta-analysis methods but by the limitations of the research literature. These limitations do not have to be permanent. Consider an example. McDaniel et al. (1988b) found that only 15 criterion-related studies had ever been conducted on the validity of the behavioral consistency method of evaluating applicants' past job-related achievements and accomplishments. Based on these 15 studies, mean true validity is estimated at .45 (SD = .10; 90% credibility value = .33; percentage variance accounted for = 82%). The appropriate interpretation of these findings is different from the interpretation that would be appropriate for exactly the same findings based on exactly the same number of studies in a meta-analysis of cognitive ability. There are literally hundreds of meta-analyses of cognitive abilities and job performance to which the latter findings could be cross-referenced to check for consistency. In the case of the behavioral consistency method, there are no other meta-analyses. Furthermore, we have very little information as to precisely what the behavioral consistency procedure measures. For example, there are no reported correlations between cognitive ability test scores and behavioral consistency scores. Behavioral consistency scores are not yet part of a rich, structured, complex, and elaborated network of established knowledge as cognitive abilities are. Therefore, this meta-analysis must stand alone to a much greater extent. We cannot be really certain that the results are not substantially influenced by outliers or by second-order sampling error. (For example, the actual amount of variance due to artifacts may be 100%, or it may be 50%.) For these reasons, McDaniel et al. (1988b) stated that these findings must be considered preliminary and recommended that additional validity studies be conducted, not to estimate "local validities" from local studies for local settings but to have more studies to combine into the meta-analysis.

There are other areas of research in industrial-organizational psychology completely outside the area of personnel selection and many areas outside the field of industrial-organizational psychology where (1) the number of studies now available is small, and (2) there is no elaborated structure of empirical and theoretical knowledge against which the meta-analytic results can be checked. When meta-analytic results have less evidentiary value because the number of individual studies in the meta-analysis is small *and* there is no related structure of empirical and theoretical knowledge against which the meta-analytic results can be

checked, the alternative is neither reversion to reliance on the single study nor a return to the narrative review method of integrating study findings; both are vastly inferior to meta-analysis in information yield. The appropriate reaction is to accept the meta-analysis provisionally while conducting (or awaiting) additional studies, which are then incorporated into a new and more informative meta-analysis. During this time, other forms of evidence bearing on the hypothesis in question may appear—forms of evidence analogous to the within-setting studies (Schmidt & Hunter, 1984; Schmidt, Ocasio, et al., 1985) in the area of situational specificity, in that they represent different approaches to the same question. Such evidence then constitutes the beginning of the kind of structured pattern of evidence described previously.

HIERARCHICAL ANALYSIS OF MODERATOR VARIABLES VIA SUBGROUPING

One approach to detection of moderators is subgrouping of studies. But the results of subgrouping can be deceptive if moderators are correlated. In searching for moderator variables using meta-analysis, some authors have used partially hierarchical subgrouping. First, all studies are included in an overall meta-analysis. The studies are then broken out by one key moderator variable, then the studies are recombined and broken out by another key moderator variable, and so on. The meta-analysis of assessment center validities by Gaugler et al. (1987) is an example of this approach. This type of analysis, however, is not fully hierarchical because the moderator variables are not considered in combination, which can result in major errors of interpretation. These errors are analogous to problems in analysis of variance due to confounding and interaction. An analysis of each moderator separately may lead to quite misleading results. In a meta-analysis by Rodgers and Hunter (1986) of the effects of management by objectives (MBO) on productivity, the initial analysis suggested two moderator variables: top-level management commitment and length of the intervention period. Their initial analysis suggested that MBO programs with the strong support of top management increased productivity by an average of 40%, while programs without the strong support of top management had little effect. Their initial analysis also suggested that studies based on an assessment period of more than 2 years showed much larger effects than studies based on less than 2 years. However, when the studies were broken down by the two moderator variables together, the effect of time virtually vanished. Most of the long-term studies were studies with strong top-management commitment, while most of the short-term studies were studies with weak top-management commitment. Thus, the apparent impact of time horizon as a moderator variable was due to the fact that it was confounded with managerial commitment. The

difficulty in conducting fully hierarchical moderator analyses in meta-analysis is often that there are too few studies to yield adequate numbers of studies in cells beyond the two-way breakout. This simply means that it is not possible to address all moderator hypotheses at that time. As more studies accumulate over time, more complete moderator analyses can be performed.

The MBO meta-analysis illustrates the potential problems of confounding between moderators, that is, "spurious" (in the language of path analysis) mean differences for one potential moderator are produced by real differences on another. Thus, confounding results from the fact that the moderators are correlated.

The second problem is potential interaction between moderator variables. Suppose two moderator variables *A* and *B* have been found to moderate effect sizes when analyzed separately, and assume the moderator variables are independent (uncorrelated) across studies. Can we then conclude that *A* and *B* always moderate effect size? We cannot. Consider an example. Suppose the mean effect size is .30 when *A* is present versus .20 when *A* is absent, and suppose the mean effect size is .30 when *B* is present versus .20 when *B* is absent. Assume that the frequency of *A* is 50% and the frequency of *B* is 50% and that *A* and *B* are independent. Then each of the four cells obtained by considering *A* and *B* together will have a 25% frequency. Consider the mean effect sizes in the following joint breakdown table:

		Moderator A		
		Present	*Absent*	*Ave.*
Moderator B	Present	.40	.20	.30
	Absent	.20	.20	.20
	Ave.	.30	.20	.25

Consider the 50% of studies in which moderator *B* is absent. Within those studies, the presence or absence of *A* does not matter; the mean effect size is .20 in either case. Thus, *A* is a moderator variable only for the studies in which *B* is present. The statement that "*A* moderates the effect of *X* on *Y*" is false for the 50% of the studies where *B* is absent. Consider the 50% of studies in which moderator *A* is absent. Within those studies, the presence or absence of *B* does not matter; the mean effect size is .20 in either case. Thus, *B* is a moderator variable only for the studies in which *A* is present. To say "*B* moderates the effect of *X* on *Y*" is false for the 50% of the studies where *A* is absent. This means that *A* and *B* are inextricably linked as moderator variables. Within the 75% of studies in which one or

the other is absent, the mean effect size is .20, regardless of whether either variable is present or absent. The only moderating effect is that studies in which both *A* and *B* are present together differ from the other studies.

There is a rule in analysis of variance that states "If there is an interaction between two or more factors in the design, then interpretation of lower order main effects or interactions may be quite erroneous." This same rule applies to interaction between moderators. If moderators have interacting effects, then the interpretation of separate effects may be erroneous.

If the hierarchical breakdown reveals moderator variables, then the overall analysis without moderator variables is likely to be misleading. If the hierarchical analysis shows that moderator variables are correlated and/or interact, then the analysis of moderator variables separately is likely to be misleading. Thus, if a hierarchical breakdown is presented, it is critical to focus the interpretation solely on the full breakdown of the data.

Consider the partially hierarchical analysis in the meta-analysis of personnel selection validities by Schmitt, Gooding, Noe, and Kirsch (1984). These researchers first pooled correlations across all predictors (biodata, tests, interviews, and more) and all criterion measures (performance ratings, tenure, advancement, etc.). They then broke the data down by predictor and criterion separately, and finally by the two together. The combinatorial breakdown showed a strong interaction between predictor and criterion variables as moderators—as had been found in past analyses. Had they based their conclusions solely on that last analysis, they would have made no error of interpretation. Unfortunately, they based some of their conclusions on the earlier global analyses. For example, they claimed that their meta-analysis yielded results at odds with the comparable meta-analysis by Hunter and Hunter (1984). However, Hunter and Hunter broke their data down by both predictor variable and criterion variable from the beginning. Thus, the only table in Schmitt et al. comparable to the analysis of Hunter and Hunter is their final table, the combinatorial breakdown. There is no contradiction between their results in that analysis and that of Hunter and Hunter. This was brought out in a side-by-side presentation in Hunter and Hirsh (1987) that showed the analyses to be in agreement.

The analysis of multiple moderator variables separately (i.e., one by one) will be correct only if one can correctly make two assumptions: One must assume that (1) the moderator variables are independent and (2) the moderator variables are additive in their effects. In the MBO analysis of Rodgers and Hunter (1986), the commitment and time moderator variables were correlated across studies. Thus, the large difference due to the commitment variable produced a "spurious" mean difference between studies of different time lengths. If the two potential moderator variables had been independent, there could have been no spurious effect for time produced by commitment. The *AB* combination example showed that interactive moderators must always be considered together to generate correct conclusions.

If a fully hierarchical analysis is presented, it is critical to base conclusions on the highest level of interaction (i.e., the full hierarchical analysis). Schmitt et al. (1984) made an error of interpretation because they went back to an analysis with confounded interactions for one of their conclusions. Finally, it is important to recognize that one often will not have enough studies to conduct a fully hierarchical moderator analysis. If the number of studies in the cells of the fully hierarchical analysis is very small, the conclusions about moderators can only be tentative. Firmer conclusions must await the accumulation of a larger number of studies.

Use of Multiple Regression in Moderator Analysis and Mixed Meta-Analysis Models

This section explores meta-regression, multilevel meta-analysis, hierarchical linear modeling (HLM), and the mixed effects (ME) meta-analysis model. As we will see, these procedures are all closely related to one another.

META-REGRESSION: ADVANTAGES AND DISADVANTAGES

In meta-regression, *r* or *d* values are regressed onto measures of potential moderator variables that have been coded as study characteristics. This procedure has been used in meta-analyses of psychotherapy outcome studies (Smith & Glass, 1977) and the effects of class size (Smith & Glass, 1980) and many other more recent meta-analyses. Glass (1977) was the first to advocate using multiple regression to identify moderator variables in meta-analysis. He recommended and used ordinary least squares (OLS) regression, but others (e.g., Hedges & Olkin, 1985) later recommend weighted least squares regression (WLS). The use of meta-regression has the advantage that it controls (at least in theory) for any potential correlations among moderator variables, hence avoiding the problems that can plague nonhierarchical subgrouping of studies in meta-analysis. It also has the advantage of being better able to deal with continuous moderators.

The same considerations that apply to other applications of regression apply to meta-regression. Unless sample sizes are sufficiently large relative to the number of variables (predictors) in the regression equation, there is a great deal of sampling error in regression weights. As a result, simulation studies have found that the multiple *R*s produced by regression weights are often less accurate in estimating population multiple *R* values than simple equal weighting of the predictor variables (Schmidt, 1971), even when there is no ex post facto selection of predictors. Under realistic conditions, with two predictors, one must have an *N* of at least 50 for regression weights to be superior to equal weights. With six predictors, *N* must be at least 100. With

8 predictors, N must be at least 150, and with 10 predictors, at least 200 (Schmidt, 1971). Keep in mind that in meta-regression, $N = k$, the number of studies. How many meta-regression studies that examine 8 potential moderators have $k = 150$ studies? Most do not. The extent of sampling error in regression weights can be illustrated by drawing multiple samples from the same realistic population, computing regression weights on each sample, and then computing the average correlation of regression weights across samples. That is, the regression weights are treated as a vector of scores whose reliability is measured by the average correlation between these vectors of weights across samples. With four predictors, it requires an N of 500 to produce a correlation of .85 (Schmidt, 1972, Table 1). For a larger number of predictors, larger Ns are required to attain this level of reliability for the regression weights. These findings apply to meta-regression as well as to other applications of regression analysis.

Meta-regression has eight serious disadvantages. The first and most serious is the potential for massive capitalization on chance resulting in inflated multiple Rs (Raudenbush, 2009), as described near the end of Chapter 2. The square of the multiple R is then falsely interpreted as the proportion of variance in r or d values explained by the "moderators." In meta-regression applications in the literature, the appropriate shrinkage formula to adjust for the inflation in the multiple R (Cattin, 1980) is almost never applied. Even if a shrinkage formula is applied, the multiple R is still inflated if there is any ex post facto selection of the potential moderators included in the meta-regression, a frequent practice. As noted in Chapter 2, some who use this procedure focus not on the multiple R but on the statistically significant regression weights. But these are also distorted by capitalization on chance if there is any ex post facto selection of the potential moderators to be included in the regression. For example, this occurs when only those potential moderators with a large or statistically significant correlation with the effect sizes are included in the meta-regression equation, a common practice. The second major problem is that statistical power is typically low, because the number of studies (k is the relevant N) is almost never large. Because of low power, the regression weights for most *real* moderators will be nonsignificant. At least they should be, given known statistical principles, yet most moderators in the literature *are* significant, which raises suspicions about capitalization on chance and/or selective reporting (see Chapter 13). The third disadvantage is susceptibility to distortion by outlier data points. This consideration exists in all applications of regression, but it is much more serious when sample sizes are small in relation to the number of predictors (Stevens, 1984). In meta-regression, the sample size (the number of studies, k) is often as small as 15, 20, or 30, and the number of potential moderators (predictors) may be as large as 5 or 10 (or more). The fourth disadvantage stems from the fact that the obtained regression weights are unstandardized (raw score) regression weights. Because of

this, they are difficult or impossible to interpret in any substantive way and any given weight cannot be compared with other regression weights in the meta-regression, as discussed in the second section of Chapter 5. For example, the size of the regression weight on any hypothesized moderator depends on how that moderator is scaled or measured. Because different moderators are measured on different scales with different *SD*s, the magnitudes of the regression weights are not comparable and their magnitudes cannot be compared to each other, so it is not apparent which are the most important moderators. This fact leads users of meta-regression to focus almost entirely on *p* values in comparing moderators, an undesirable emphasis; *p* values become an inappropriate index of importance.

The fifth disadvantage is the fact that meta-regression results are inaccurate when the *d* or *r* values have not been corrected for measurement error (and for range restriction, where applicable) (Ones, Viswesvaran, & Schmidt, 2012). While all the values will be biased downward by measurement error, some will be biased more than others, undercutting the construct validity of the observed *d* s or *r* s as measures of the real effects and, hence, artifactually reducing the apparent strength of all true moderators. It is rare in published meta-regressions for these corrections to be made. The correction for measurement error that is necessary for accuracy of meta-regression results causes significance tests, standard errors, and confidence intervals for regression coefficients to be inaccurate with most computer programs. Hunter (1995) developed special software that yields accurate standard error values when the data have been corrected for measurement error. (Corrections to individual *r* or *d* values are not possible when artifact distribution meta-analysis is used, which makes the use of meta-regression even more questionable in such cases.) The sixth problem is measurement error in the measures of the hypothesized moderator variables. As noted in Chapter 3, Orwin and Cordray (1985) showed that failure to correct for measurement error in the measures of the moderator variables leads to serious errors in the meta-regression results. This finding is important because almost no meta-analyses in the literature using meta-regression make this correction, meaning their moderator results are suspect (see also Cordray & Morphy, 2009). Seventh, even if the *d* or *r* values are corrected for measurement error and other artifacts, there is still the problem created by the fact that much (often most, sometimes all) of the variance in the *d* values or *r*s (i.e., the dependent variable) is due to sampling error and other artifacts, creating low reliability for the dependent variable and, hence, low statistical power to detect moderator effects (as discussed in Chapters 2, 3, and 7; Cook et al., pp. 325–326). (Aloe, Becker, and Pigott, 2010, have proposed an adjustment for the sampling error in the effect sizes [which functions as measurement error in meta-regression]. This adjustment partials the effects of sampling error out of the multiple correlation and is similar to the correction for effect size unreliability illustrated in Chapter 3 in the Tibetan Employment Service example.) The

eighth disadvantage stems from problematic data requirements. Use of meta-regression requires estimates of the correlations among the potential moderators (predictors). Often, estimates of some of these correlations are not available and must be guessed at or somehow imputed. This potential adds additional error to the meta-regression results.

In articles and textbooks, meta-regression is often presented and described without any mention of these disadvantages (e.g., Lipsey & Wilson, 2001). In light of these serious limitations, it can be seen that it is only under rare and unusual circumstances that meta-regression will produce reliable and valid results. Meta-regression is used quite frequently in the literature today, and it is highly likely that most of the results are not trustworthy.

Hedges and Olkin (1985, pp. 11–12, 167–169) argued for use of weighted least squares (WLS) regression rather than OLS in meta-regression. They pointed out that the assumption of homogeneity of sampling error variances is usually not met in meta-analysis data sets. The sampling error variance of each "observation" (i.e., each *d* or *r* value) depends on the sample size on which it is based (and on the size of the observed *d* or *r* value). If these sample sizes vary substantially, as they usually do, then different effect size estimates will have different sampling error variances. In meta-regression, study sampling error variance plays the same role as measurement error in a primary study analysis. Heterogeneity of variances can affect the validity of significance tests; actual alpha levels may be larger than nominal levels (e.g., .10 vs. the nominal .05). Estimates of the regression weights of moderators and multiple correlations can also be affected. Hedges and Olkin (1985, chap. 8) described a WLS regression procedure that circumvents these potential problems by weighting each study by the inverse of its sampling error variance. However, when Hedges and Stock (1983) used this method to reanalyze the Smith and Glass (1980) studies on class size, they obtained results that were quite similar to the original results, suggesting that the problem identified by Hedges and Olkin (1985) may not be serious when the number of *d* or *r* values is large (which was the case in the Smith and Glass, 1980, study). The general finding has been that most statistical tests are robust with respect to violations of the assumption of homogeneity of variance (see, e.g., Glass, Peckham, & Sanders, 1972; or Kirk, 1995).

In an attempt to address this question, Steel and Kammeyer-Mueller (2002) compared OLS and WLS using computer simulation. They focused only on continuous moderators and only on the accuracy of the multiple *R* resulting from predicting observed effect sizes from the continuous moderator variables. They did not look at the accuracy of the standardized regression weights, which provide the needed information on the size and importance of each individual moderator variable. They found that when the distribution of study sample sizes (*N*) was approximately normal, there was little difference in the accuracy of OLS and WLS. However, when the distribution of study *N*s was skewed to the right, WLS produced more accurate estimates of the multiple *R*. However, the level of skew they examined was somewhat

extreme (skew = 2.66) and might occur only infrequently in real research literatures. The Steel and Kammeyer-Mueller study did not address or discuss the problem of capitalization on sampling error in the use of either type of regression weighting. Nor did the study address the other disadvantages of meta-regression discussed above.

There are reasons to be cautious about the use of WLS. If there is an outlier (in either direction) with an extremely large *N*, the WLS estimates will be greatly influenced by such a study. Hence, with WLS, it is especially important to be concerned with outliers. There are also potential problems in the weighting of studies with small *N* s. When *N* is small, very large *r* or *d* values can occur due to large sampling errors. The observed value of *r* or *d* affects the computed sampling error variance (as can be seen by inspecting the formulas for sampling error variance for *r* and *d*) and, therefore, affects the weight the study gets. As Steel and Kammeyer-Mueller (2002) noted, a study based on $N = 20$ with an *r* of .99 would be given the same weight as a study based on $N = 20{,}000$ but with an *r* of .60! One solution to this latter problem is to use mean *r* or *d* in the sampling error variance formulas for *r* and *d*, in place of the *r* and *d* values from the individual study, as discussed and recommended in Chapters 3, 4, 5, and 7. Because both OLS and WLS regression methods have (different) problems, Overton (1998) recommended applying both and comparing the results. If they are similar, one's confidence in the results is supported. However, the eight disadvantages of meta-regression discussed above remain whether WLS or OLS is used. Modification of the study weights does not make these problems go away.

With these cautions in mind, when moderators are continuous and the decision has been made to use regression, it is probably advisable in typical cases to emphasize WLS results in preference to OLS results. In most meta-analyses that use meta-regression, the meta-regression analysis is conducted after the main meta-analysis. However, some applications of meta-analysis consist of only a meta-regression analysis. In general, this is not an approach that we recommend because it does not produce an overall corrected mean and standard deviation for the population parameters. In addition, the results produced by this approach will be stable only if *k*, the number of studies, is very large. The meta-analysis by Nye, Su, Rounds, and Drasgow (2012) used this approach effectively. This meta-analysis included 568 correlations, so sampling error in the regression analysis was greatly reduced. Such large *k* values are rare, however. An example of this approach to meta-analysis is the structural equation modeling (SEM)–based meta-analysis methods of M. W. L. Cheung (2008), discussed in Chapter 11. In practice, this form of meta-analysis, including Cheung's approach, is usually a mixed effects meta-analysis (discussed later). This approach to meta-analysis can be viewed as a form of hierarchical linear modeling (discussed later).

If moderators are dichotomous or categorical (e.g., sex or race), the subgrouping approach to moderator analysis is superior. However, it is important to bear in mind that moderators are often correlated and that it is important to use hierarchical moderator analysis to avoid confounding of correlated moderators. When moderators are continuous, the subgrouping method has the disadvantage of requiring dichotomization of the continuous variables to produce the subgroups, thus losing information. When there is only one hypothesized moderator to be examined and it is continuous, simple correlation can be used, as described in Chapters 3 and 7. That correlation is then the standardized regression weight for predicting the effect sizes or correlations. In this case, there is no capitalization on chance. When there is more than one continuous moderator, simple correlation is maximally informative only if the moderators are uncorrelated. If the continuous moderators are correlated, OLS or WLS can be used to assess the moderators (bearing in mind the limitations of meta-regression). Hierarchical meta-analysis via subgrouping can also be used, but it requires dichotomizing (or perhaps trichotomizing) the continuous moderator variables, which is not desirable. It is not clear which of these two options is to be preferred in a case like this. Some advice on use of meta-regression in moderator detection is provided by Aguinis and Gottfredson (2010) and Aguinis and Pierce (1998).

MULTILEVEL MODELS IN META-ANALYSIS AND HLM

The use of meta-regression as described in the previous section is often referred to as "multilevel" meta-analysis. In this nomenclature, the first level is the meta-analysis of *d* or *r* values, and the second level is the regression of the effect sizes onto a set of potential moderator variables. In a sense, this is a form of hierarchical linear modeling (HLM; Raudenbush, 2009; Raudenbush & Bryk, 2002). But HLM is typically used when effect sizes are not independent. For example, in educational research, teachers are nested within classrooms, and therefore the academic achievement scores of students within the same classroom are not independent (Raudenbush & Bryk, 2002), because the achievement of all the students is affected by the competence of their particular teacher. In meta-analysis, HLM is typically restricted to the case in which the same sample or study contributes multiple *r* or *d* values, which creates a similar violation of the assumption of independence. HLM can handle this problem. Freund and Kasten (2012) is an example of such an application of HLM in a meta-analysis. However, as seen in Chapter 10, we recommend that steps be taken to ensure that the effect sizes within a meta-analysis are statistically independent of each other. We also present evidence that violations of the independence assumption have less of a distorting effect on meta-analysis results than is usually assumed. It has been suggested that the HLM can be viewed

as a general approach to conducting meta-analyses (Raudenbush & Bryk, 2002). As such, it is a form of linear mixed effects meta-analysis (discussed in the next section). However, in practice, HLM has typically been restricted to cases in which the independence assumption has been violated. Hedges, Tipton, and Johnson (2010a, 2010b) have presented an approach to HLM that is simpler to use and is robust to the distributional assumptions made by other HLM approaches. A major limitation of HLM in general is that it is very difficult, if not impossible, to correct for the distorting effects of measurement error and range restriction or enhancement, causing inaccuracy in the results. HLM has all of the eight disadvantages of meta-regression discussed in the previous section. In addition, most applications of HLM use maximum likelihood (ML) estimation methods, which require larger sample sizes. As in the case of meta-regression, it is only under rare circumstances that one has data sufficient to cause HLM to produce accurate results.

MIXED EFFECTS MODELS IN META-ANALYSIS

In Chapters 5 and 8, we presented discussions of fixed effects (FE) and random effects (RE) models in meta-analysis. There is a related concept called the "mixed effects (ME) model." The ME model is viewed as a mixture of RE and FE models. Suppose a meta-analyst applies the RE model to a set of effect sizes and stops after calibrating the variation in the population effect sizes, with no attempt to test or identify moderators. This occurs when the meta-analyst views this variation as completely random; that is, produced by unknown (and maybe unknowable) factors. This is referred to by Hedges (1982c, 1983b) as the "simple random effects model." Alternatively, the meta-analyst might hypothesize that certain specific factors account for at least some of the between-study variability in population values. The meta-analyst would then attempt to test these hypotheses using meta-regression (Raudenbush, 2009). According to Hedges (1983b), if these hypothesized moderators are related to study outcomes, they are then viewed as "fixed factors," meaning that they constitute all the potential moderators that the researcher is or would ever be interested in. This is the definition of fixed factors in analysis of variance (National Research Council, 1992). Therefore, in the weighted meta-regression, the studies are weighted by the inverse of their FE sampling error variances and not by the inverse of their larger RE sampling error variances (Overton, 1998; Raudenbush, 2009), which would be the study weights used if the moderators were viewed as just a sample of possible moderators. If there is no further variation in population effect sizes beyond sampling error in the subgroups in which these moderators are held constant (i.e., if the hypothesized moderators produce a [properly adjusted] multiple R of 1.00 with the actual study effect sizes), then the overall model is said to be an FE

model, because there is no unexplained variation left. This outcome is rare in real data, if it exists at all. If, on the other hand, the postulated moderators account for some but not all the variation in the rho or delta values, the resulting model is said to be an ME model. The RE part of this conclusion stems from the fact that there is still remaining unexplained variance in population parameters, variance not accounted for by the moderators. The FE part of this conclusion stems from the fact that the postulated moderator variables are assumed to be fixed factors (FE factors). Vevea and Citkowicz (2008) showed via simulation that this approach often results in "seriously inflated Type I errors" in testing potential moderators that are in fact unrelated to the effect sizes. But in addition to this statistical problem there is also a conceptual problem here. If variance in population parameters remains after controlling for the fixed factors, there must be other moderators operating. This fact casts doubt on the definition of the fixed factors as constituting all the moderators that the researcher is interested in or could ever be interested in (National Research Council, 1992). The SEM-based methods of M. W. L. Cheung (2008) are ME meta-analysis methods, as noted earlier.

This way of thinking about meta-analysis models stemmed from an early conception of meta-analysis models in the 1985 Hedges and Olkin meta-analysis book. At that time, the hope was that the FE model as described above would turn out to be the case. That is, the hope was that postulated moderator variables would account for all the variation beyond sampling error in study population values. If so, then the FE model as defined above would actually apply. However, as meta-analyses accumulated in the literature, it became apparent that postulated moderators almost never explained all the variance in population parameters. (This result could be due in part to the fact that these meta-analysis methods do not control for variation due to artifacts such as measurement error; see Chapter 11. It is possible that these artifacts account for the remaining variance.)

Second-Order Sampling Error: General Principles

The outcome of any meta-analysis based on a small number of studies depends to some extent on which studies randomly happen to be available; that is, the outcome depends in part on study properties that vary randomly across studies. This is true even if the studies analyzed are all that exist at that moment. This phenomenon is called "second-order sampling error." It affects meta-analytic estimates of standard deviations more than it affects estimates of means. This is also the case with ordinary, or first-order, sampling error and ordinary statistics: Ordinary sampling error affects standard deviations more than means. Ordinary, or first-order, sampling error stems from the sampling of subjects within a study. Second-order sampling error stems from the sampling of studies in a meta-analysis.

Consider a hypothetical example. Suppose there were only 10 studies available estimating the relationship between Trait A and job performance. Even if the mean sample size per study were only 68 (the median for published validity studies reviewed by Lent et al., 1971a, 1971b), the mean validity would be based on $N = 680$ and would be reasonably stable. The observed variance across studies would be based on only 10 studies, however, and this variance, which we compare to the amount of variance expected from sampling error, would be based on only 10 data points. Now suppose sampling error were, in fact, the only factor operating to produce between-study variance in observed correlations (validities). Then, if we randomly happened to have one or two studies with large positive sampling errors, the observed variance across studies would likely be larger than the variance predicted by the sampling error variance formula, and we might falsely wind up concluding, for example, that sampling error accounts for only 50% of the observed variance of validities across studies. On the other hand, if the observed validity coefficients of, say, five or six of the studies randomly happened to be very close to the expected value (population mean), then the observed variance across studies would likely be very small and would underestimate the amount of variance one would typically (or on the average) observe across 10 such randomly drawn studies (from the population of such hypothetical studies that could be conducted). In fact, the observed variance might be smaller than the variance predicted from sampling error. The computed percentage variance accounted for by sampling error would then be some figure greater than 100%, for example, 150%. Of course, in this case, the correct conclusion would be reached: All the observed variance could be accounted for by sampling error. However, some people have been troubled by such outcomes. They are taken aback by results indicating that sampling error can account for more variance than is actually observed. Sometimes they are led to question the validity of the formula for sampling error variance (see, e.g., H. Thomas, 1988, and the reply by Osburn & Callender, 1990). This formula correctly predicts the amount of variance sampling error will produce on average. However, sampling error randomly produces more than this amount in some samples and less in other samples. The larger the number of studies (other things being equal), the smaller the deviations of observed from expected sampling variance. If the number of studies is small, however, these deviations can be quite large *on a percentage basis* (although *absolute* deviations are usually small even in such cases).

Negative estimates of variances occur using other methods of statistical estimation. In one-way analysis of variance (ANOVA), for example, the variance of sample means is the sum of two components: the variance of population means and the sampling error variance. This is directly analogous to the meta-analytic breakdown of the observed variance of sample correlations across studies into the variance of population correlations (the real variance) and the sampling error variance (the false

or spurious variance). In estimating the variance of population means in ANOVA, the first step is to subtract the within-group mean square from the between-group mean square. This difference can be, and sometimes is, negative, as a result of sampling error. Consider a case in which the null hypothesis is true; the population means are then all equal and the variance of population means is 0. The variance of observed means (i.e., sample means) is then determined entirely by sampling error. This observed between-group variance will vary randomly from one study to another. About half the time, the within-group mean square will be larger than the between-group mean square, while half the time, the within-group mean square will be smaller. That is, if the variance of population means is 0, then in half of the observed samples, the estimated variance of population means will be negative. This is exactly the same as the situation in meta-analysis if all the population correlations are equal: The estimated variance will lie just above 0 half the time and will lie just below 0 half the time. The key here is to note that the variance of population correlations is estimated by subtraction: The known sampling error variance is subtracted from the variance of sample correlations, which estimates the variance of sample correlations across a population of studies. Because the number of studies is never infinite, the observed variance of sample correlations will depart by sampling error from the expected value. Thus, when the variance of population correlations is 0, the difference will be negative half the time.

Another example is the estimation of variance components in generalizability theory. Cronbach and his colleagues (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) proposed generalizability theory as a liberalization of classical reliability theory, and it is now widely used to assess the reliability of measuring instruments in situations where the techniques of classical reliability theory are considered inadequate. Generalizability theory is based on the well-known ANOVA model and requires estimated variance components for its application. One or more of the estimated variance components may be negative, as noted by Cronbach et al. (1972, pp. 57–58) and Brennan (1983, pp. 47–48), even though, by definition, population variance components are nonnegative. The same phenomenon was also noted by Leone and Nelson (1966). Cronbach et al. (1972) recommended substituting 0 for the negative variance, and Brennan (1983) agreed with this recommendation.

Negative estimated variances are not uncommon in statistical estimation. The occurrence of negative estimates of variance in empirical research does not call into question a statistical theory such as ANOVA or a psychometric theory such as meta-analysis. As described previously, existing statistical sampling theory provides a sound rationale for observed negative estimates of variance in meta-analysis when the actual variance of true validities is 0 or close to 0. W. A. Thompson (1962) provides an analytical discussion of negative variance estimates.

Second-Order Meta-Analyses Across Different Independent Variables

A second-order meta-analysis is a meta-analysis of meta-analyses. A form of second-order meta-analysis can be applied in cases in which the different meta-analyses have nonidentical in dependent variables to which the same theoretical and methodological considerations apply. In such cases, the effect sizes cannot be combined because the in dependent variables are different, but the problem of second-order sampling error can be addressed by computing the average percent variance accounted for by artifacts. Validity generalization research on cognitive ability tests is an example. Under the situational specificity hypothesis, the hypothesized situational moderators would be essentially the same for different abilities (e.g., verbal, quantitative, reasoning, spatial), and under the alternate hypothesis, all variance would be hypothesized to be artifactual for all abilities. The second-order meta-analysis would involve computing the average percentage of variance accounted for across the several meta-analyses. For example, in a large consortium study conducted by Psychological Services, Inc., in 16 companies, the percentage of variance accounted for by sampling error ranged over different abilities from about 60% to more than 100%. The average percentage accounted for across abilities was 99%, indicating that once second-order sampling error was considered, all variance of validities across the 16 companies was accounted for by sampling error for all the abilities studied.

Such a finding indicates that the meta-analyses with less than 100% of the observed variance accounted for are explained as cases of second-order sampling error (specifically, secondary second-order sampling error, as defined later in this chapter). The same is true of meta-analyses with more than 100% of the observed variance accounted for. It should be clear that in conducting a second-order meta-analysis, figures greater than 100% should not be rounded down to 100%. Doing so would obviously bias the mean for these figures, because those that are randomly lower than 100% are not rounded upward.

There is an important technical issue in this form of second-order meta-analysis: The average percentage of variance accounted for must be computed in a particular way or it will be inaccurate. This technical issue is best illustrated by a study conducted by Spector and Levine (1987). Spector and Levine conducted a computer simulation study aimed at evaluating the accuracy of the formula for the sampling error variance of *r*. In their study, the value of ρ was always 0, so the formula for the sampling error variance of observed *r*s was $S_e^2 = 1/(N-1)$. They conducted simulation studies for various values of *N*, ranging from 30 to 500. The number of observed *r*s per meta-analysis was varied from 6 to 100. For each combination of *N* and number of *r*s, they replicated the meta-analysis 1,000

times and then evaluated the average value of S_e^2 / S_r^2 across 1,000 meta-analyses. That is, they focused their attention on the average ratio of variance predicted from the sampling error formula to the average observed variance of the *r* s across studies. They did not look at $S_r^2 - S_e^2$, the difference between predicted and observed variances. They found that for all numbers of *r*s less than 100, the ratio S_e^2 / S_r^2 averaged more than 1.00. For example, when there were 10 *r*s per meta-analysis and $N = 75$ in each study, the average ratio was 1.25. Kemery, Mossholder, and Roth (1987) obtained similar results in their simulation study. The smaller the number of *r*s per meta-analysis, the more the ratio exceeded 1.00. They interpreted these figures as demonstrating that the formula for S_e^2 overestimates sampling variance when the number of correlations in a meta-analysis is less than 100. Their assumption was that if the S_e^2 formula were accurate, the ratio S_e^2 / S_r^2 would average 1.00.

The Spector-Levine (1987) study was critiqued by Callender and Osburn (1988), who showed that if one assessed accuracy by the difference $S_r^2 - S_e^2$, the sampling error variance formula was shown to be extremely accurate, as had also been demonstrated in their numerous previous simulation studies. There was no bias. They also demonstrated why the average ratio S_e^2 / S_r^2 is greater than 1.00 despite the fact that S_e^2 is an unbiased estimate of sampling variance. When the number of correlations in a meta-analysis is small, then, by chance, the S_r^2 will sometimes be very small; that is, by chance, all observed *r*s will be very similar to each other. Because S_r^2 is the denominator of the ratio, these tiny S_r^2 values lead to very large values for S_e^2 / S_r^2, sometimes as large as 30 or more. Furthermore, if S_r^2 should, by chance, be 0, the ratio is *infinitely large*. These extreme values raise the mean ratio above 1.00; the *median* ratio is very close to 1.00. The analysis by Callender and Osburn (1988) fully explains the startling conclusions of Spector and Levine (1987) and demonstrates that the fundamental sampling variance formula for the correlation is, in fact, accurate.

It should be noted that Spector and Levine would not have reached the conclusion they reached had they used the reciprocal of their ratio. That is, if they had used S_r^2 / S_e^2 instead of S_e^2 / S_r^2, they would have found that the mean ratio was 1.00. With this reversed ratio, the most extreme possible value is 0 (rather than infinity), and the distribution of ratios is much less skewed. This point has important implications for second-order meta-analyses of the sort discussed in this section. As noted earlier, this form of second-order meta-analysis is conducted by averaging the percentage of variance accounted for by artifacts over similar meta-analyses. In any given meta-analysis, this percentage is the ratio of artifact-predicted variance (sampling variance plus variance due to other artifacts) to the

observed variance. One over this ratio is the reversed ratio, S_r^2 / S_e^2. In second-order meta-analysis, this reversed ratio should be averaged across studies, and then the reciprocal of that average should be taken. This procedure prevents the upward bias that appeared in the Spector-Levine study and results in an unbiased estimate of the average percentage of variance across the meta-analyses that is due to artifacts. For an example application, see Rothstein et al. (1990).

Meta-analysis has made clear how little information there is in single studies because of the distorting effect of (first-order) sampling error. An examination of second-order sampling error shows that even several studies combined meta-analytically contain limited information about between-study variance (although they provide substantial information about means). Accurate analysis of between-study variance requires either meta-analyses based on a substantial number of studies (we have had up to 882; cf. Pearlman et al., 1980) or meta-analyses of similar meta-analyses (second-order meta-analyses). These are the realities and inherent uncertainties of small-sample research in the behavioral and social sciences (or in any other area, e.g., the biomedical area). There is no perfect solution to these problems, but meta-analysis is the best available solution. As the number of studies increases, successive meta-analyses will become increasingly more accurate.

Second-Order Meta-Analysis With a Constant Independent Variable

When there are number of independent meta-analyses that focus on the same independent and dependent variables, another form of second-order meta-analysis becomes possible (Schmidt & Oh, 2013). For example, Oh (2009) conducted separate meta-analyses of the validity of five personality traits for predicting job performance in four East Asian countries. Each meta-analysis contained only studies conducted in that country, so there were no overlapping studies between meta-analyses. Mean validity values differed across countries but a second-order meta-analysis using the methods described in this section showed that most of the between-country variance in mean correlation values was due to second-order sampling error. For one personality trait—Conscientiousness—all the between-country variability was due to second-order sampling error, indicating that this trait has the same mean validity in all the countries. Schmidt and Oh (2013) present other such applications. The number of meta-analyses of the same independent and dependent variables conducted in different countries or regions is increasing, and so this form of second-order meta-analysis is becoming more important.

In this section, we present the essential equations and computations for second-order meta-analysis applied to (a) bare-bones meta-analyses (as described in Chapter 3), (b) meta-analyses that corrected each value individually (as described in Chapter 3), and (c) meta-analyses that used the

artifact distribution method to correct for artifacts (as described in Chapter 4). The presentation is in terms of correlations, but analogous equations apply when the outcome statistic is the *d* value.

SECOND-ORDER META-ANALYSIS OF BARE-BONES META-ANALYSES

Equation (9.1) is the fundamental equation when the first-order meta-analyses entering the second-order meta-analysis have corrected only for sampling error:

$$\hat{\sigma}^2_{\bar{\rho}_{xy}} = S^2_{\hat{\bar{r}}} - E(S^2_{e_{\hat{\bar{r}}_i}}) \tag{9.1}$$

where the term on the left side of the equation is the estimate of the population variance of the uncorrected mean correlations ($\hat{\bar{\rho}}_{xy}$) across the meta-analyses after second-order sampling error has been subtracted out. The first term on the right side of Equation (9.1) is the weighted variance of the mean correlations across the *m* meta-analyses, computed as follows:

$$S^2_{\hat{\bar{r}}} = \sum_1^m w_i \left(\hat{\bar{r}}_i - \hat{\bar{\bar{r}}} \right)^2 / \sum_1^m w_i \tag{9.1a}$$

where

$$\hat{\bar{\bar{r}}} = \sum_1^m w_i \hat{\bar{r}}_i / \sum_1^m w_i \tag{9.1b}$$

and

$$w_i = \left(\frac{S^2_{r_i}}{k_i} \right)^{-1} \tag{9.1c}$$

and where $S^2_{r_i}$ is the variance of the observed correlations (*r*s) in the *i*th meta-analysis, $\hat{\bar{r}}_i$ is the estimate of the mean effect size for the *i*th meta-analysis, $\hat{\bar{\bar{r}}}$ is the estimate of the (weighted) grand mean effect size across the *m* meta-analyses, k_i is the number of primary studies included in the *i*th meta-analysis, and the w_i is the weight applied to the *i*th meta-analysis. The second term on the right side of Equation (9.1) is the expected (weighted average) second-order sampling error variance across the *m* meta-analyses:

$$E(S^2_{e_{\hat{\bar{r}}_i}}) = \sum_1^m \left(w_i \frac{S^2_{r_i}}{k_i} \right) / \sum_1^m w_i \tag{9.1d}$$

Equation (9.1d) reduces to Equation (9.1e):

$$E(S^2_{e_{\hat{\bar{r}}_i}}) = m / \sum_1^m w_i \tag{9.1e}$$

To summarize, each meta-analysis will have reported a mean uncorrected (i.e., mean observed) correlation, $\hat{\bar{r}}_i$. The first term on the right in Equation (9.1) is the weighted variance of these mean correlations. This computation is shown in Equations (9.1a) and (9.1b). The weights (w_i) used in Equations (9.1a), (9.1b), (9.1d), and (9.1e) are as defined in Equation (9.1c). Each weight is the inverse of the random effect (RE) sampling error variance for the mean correlation in the *i*th meta-analysis (Schmidt, Oh, & Hayes, 2009). The second term on the right in Equation (9.1) is the sampling error variance of these mean correlations. Each of the meta-analyses will have reported the variance of the observed correlations in that meta-analysis. Dividing each such variance by k_i (the number of studies in that meta-analysis) yields the RE sampling error variance of the mean *r* ($\hat{\bar{r}}_i$) in that meta-analysis (Schmidt, Oh, & Hayes, 2009). (This reflects the well-known principle that the sampling error variance of the mean of any set of scores is the variance of the scores divided by the number of scores [and the standard error of the mean is the square root of this value.]) The weighted average of these values across the *m* meta-analyses estimates the RE sampling error variance of the mean *r*s as a group, as shown in Equations (9.1d) and (9.1e). The square root of this value divided by the square root of *m* is the standard error ($SE_{\hat{\bar{\bar{r}}}}$) and can be used to put confidence intervals around the estimate of the (weighted) grand mean ($\hat{\bar{\bar{r}}}$; computed in Equation [9.1b]). Also, using the square root of the value on the left side of Equation 9.1 ($\hat{\sigma}_{\bar{\rho}_{xy}}$), one can construct a credibility interval (see Chapters 5 and 8) around the grand mean correlation across the *m* meta-analyses, within which a given percentage of the first-order population meta-analytic (mean) effect sizes ($\hat{\bar{\rho}}_{xy}$) is expected to lie. For example, 80% would be expected to lie within the 80% credibility interval. If the value on the left side of Equation (9.1) is zero, the conclusion is that the mean population correlation values are the same across the meta-analyses. In that case, all the observed variance is accounted for by second-order sampling error, and the conclusion is that there are no moderators. If it is greater than zero, one can compute the proportion of variance between meta-analyses that is due to

second-order sampling error. This is computed as the ratio of the second term on the right side of Equation (9.1) to the first term on the right side, that is,

$$\text{ProportionVar} = E(S^2_{e_{\hat{\bar{r}}_i}}) / S^2_{\hat{\bar{r}}} \tag{9.1f}$$

and 1 − ProportionVar denotes the proportion of the variance across first-order meta-analytic (bare-bones) mean correlations that is "true" variance (i.e., variance not due to second-order sampling error). As such, this number is the reliability of the meta-analytic correlations (considered as a set of values, one for each first-order meta-analysis; see Chapters 3 and 7). This follows because reliability is the proportion of total variance that is true variance (Magnusson, 1966; Nunally & Bernstein, 1994). As discussed later, this value can be used to produce enhanced accuracy for estimates of these mean (meta-analytic) correlations from the first-order meta-analyses by regressing them toward the value of the grand mean correlation (the mean across the first-order meta-analyses). Both of these analyses are unique to second-order meta-analysis and cannot be performed using other analysis methods.

SECOND-ORDER META-ANALYSIS WHEN CORRELATIONS HAVE BEEN INDIVIDUALLY CORRECTED

Measurement error is present in all research, and it biases all relationships examined in research. Therefore, it is important to include corrections for these biases. One approach in meta-analysis is to correct each correlation individually for the downward bias created by measurement error and other artifacts as appropriate (see Chapter 3). When the first-order meta-analyses entering the second-order meta-analysis have corrected each correlation individually for measurement error (and range restriction and dichotomization, if applicable), the fundamental equation for second-order meta-analysis is

$$\hat{\sigma}^2_{\bar{\rho}} = S^2_{\hat{\bar{\rho}}} - E(S^2_{e_{\hat{\bar{\rho}}_i}}) \tag{9.2}$$

where the term on the left in Equation (9.2) is the estimate of the actual (nonartifactual) variance across the m meta-analyses of the population mean disattenuated correlations ($\hat{\bar{\rho}}$), that is, the variance after variance due to second-order sampling error has been subtracted out. The first term on the right side of Equation (9.2) is the variance of the mean individually corrected correlations across the m meta-analyses, computed as follows:

$$S^2_{\hat{\bar{\rho}}} = \sum_1^m w_i^* \left(\hat{\bar{\rho}}_i - \hat{\bar{\bar{\rho}}} \right)^2 / \sum_1^m w_i^* \tag{9.2a}$$

where

$$\hat{\bar{\bar{\rho}}} = \sum_1^m w_i^* \hat{\bar{\rho}}_i / \sum_1^m w_i^* \tag{9.2b}$$

and

$$w_i^* = \left(\frac{S^2_{r_{c_i}}}{k_i} \right)^{-1} \tag{9.2c}$$

and where $S^2_{r_{c_i}}$ is the weighted variance of the disattenuated (individually corrected) correlations in the *i*th meta-analysis, $\hat{\bar{\rho}}$ is the mean meta-analytic disattenuated correlation in that meta-analysis, $\hat{\bar{\bar{\rho}}}$ is the (weighted) grand mean effect size across the *m* meta-analyses, k_i is the number of primary studies included in the *i*th meta-analysis, and the w_i^* is the weight applied to the *i*th meta-analysis. The second term on the right side of Equation (9.2) is the weighted average second-order sampling error variance across the *m* meta-analyses:

$$E(S^2_{e_{\hat{\bar{\rho}}_i}}) = \sum_1^m w_i^* \left(\frac{S^2_{r_{c_i}}}{k_i} \right) / \sum_1^m w_i^* \tag{9.2d}$$

Equation (9.2d) reduces to Equation (9.2e):

$$E(S^2_{e_{\hat{\bar{\rho}}_i}}) = m / \sum_1^m w_i^* \tag{9.2e}$$

where the w_i^* are as defined in Equation (9.2c).

To summarize, each first-order meta-analysis will have reported an estimate of the mean disattenuated correlation (the meta-analytic mean correlation, $\hat{\bar{\rho}}_i$). The first term on the right side of Equation (9.2) is the variance of these meta-analytic mean correlations across first-order meta-analyses. This computation is shown in Equations (9.2a) and (9.2b). Equation (9.2c) shows the weights that are used in Equations (9.2a) and (9.2b). The second term on the right side of Equation (9.2) is the expected value of the second-order sampling error variance of these meta-analytic correlations. Each meta-analysis will have reported an estimate of the variance of the corrected correlations it included, preferably to four decimal places, for precision. Dividing this value by *k* (the number of studies in the

meta-analysis) yields the RE sampling error variance of the meta-analytic correlation for that meta-analysis (Schmidt, Oh, & Hayes, 2009). (As noted earlier, this reflects the well-known statistical principle that the sampling error variance of the mean of any set of scores is the variance of the scores divided by the number of scores [and the standard error of the mean is the square root of this value].) As shown in Equations (9.2d) and (9.2e), the weighted mean of these values across the *m* meta-analyses yields the second-order sampling error variance needed in Equation (9.2).

The square root of this value divided by the square root of *m* is the standard error ($SE_{\bar{\hat{\rho}}}$) and can be used to put confidence intervals around the grand mean ($\bar{\hat{\rho}}$; shown in Equation [9.2b]).

The term on the left side of Equation (9.2) is the estimate of the actual (nonartifactual) variance across meta-analysis of the population mean disattenuated correlations (the $\hat{\bar{\rho}}_i$), that is, the variance across first-order meta-analytic estimates after removal of variance due to second-order sampling error. Using the square root of this value ($\hat{\sigma}_{\bar{\rho}}$), credibility intervals can be placed around the grand mean computed in Equation (9.2b).

If the value on the left side of Equation (9.2) is zero, the indicated conclusion is that the mean population correlation values are the same across the multiple meta-analyses. All the variance is accounted for by second-order sampling error. If this value is greater than zero, one can compute the proportion of variance across meta-analyses that is explained by second-order sampling error. This is computed as a ratio of the second term on the right side of Equation (9.2) to the first term on the right side, that is,

$$\text{ProportionVar} = E(S^2_{e_{\hat{\bar{\rho}}_i}}) / S^2_{\hat{\bar{\rho}}} \tag{9.2f}$$

and 1 – ProportionVar denotes the proportion of the variance across the first-order meta-analysis mean population correlation values that is true variance (i.e., variance not due to second-order sampling error). As such, this number is the reliability of the estimated mean first-order population correlations (see Chapter 3), because reliability is the proportion of total variance that is true variance (Magnuson, 1966; Nunnally & Bernstein, 1994). This value can be used to refine the estimates of these first-order meta-analysis mean values by regressing them toward the value of the grand mean disattenuated correlation (the mean across the *m* meta-analyses, computed in Equation [9.2b]). In addition, when $S^2_{\hat{\bar{\rho}}}$ is zero, the ProportionVar is 100% and the reliability of the vector of *m* first-order meta-analytic mean estimates is zero (e.g., Conscientiousness in Table 2 of Schmidt & Oh, 2013). This is the same as the situation in which all examinees get the same score on a test, making the reliability of the test zero.

SECOND-ORDER META-ANALYSIS WITH ARTIFACT DISTRIBUTION META-ANALYSES

Often the information needed to correct each correlation individually for measurement error is unavailable for many or most of the studies. In such literatures, meta-analysis can nevertheless correct for measurement error by use of measurement error estimates (reliability estimates) from other credible sources, as indicated earlier. This method of meta-analysis is called artifact distribution meta-analysis (see Chapter 4). Equation (9.3) is the fundamental equation for second-order meta-analysis when the first-order meta-analyses have applied the artifact distribution method of meta-analysis.

$$\hat{\sigma}^2_{\bar{\rho}} = S^2_{\hat{\bar{\rho}}} - E(S^2_{e\,\hat{\bar{\rho}}_i}) \tag{9.3}$$

where the term on the left side of Equation (9.3) is the estimate of the nonartifactual variance of the population meta-analytic (disattenuated) correlations (population parameter values) across the *m* first-order meta-analyses. This is the variance remaining after variance due to second-order sampling error has been subtracted out. The first term on the right side of Equation (9.3) is the variance of the mean disattenuated correlations across the *m* meta-analyses, computed as follows:

$$S^2_{\hat{\bar{\rho}}} = \sum_1^m w_i^{**} \left(\hat{\bar{\rho}}_i - \hat{\bar{\bar{\rho}}} \right)^2 / \sum_1^m w_i^{**} \tag{9.3a}$$

where

$$\hat{\bar{\bar{\rho}}} = \sum_1^m w_i^{**} \hat{\bar{\rho}}_i / \sum_1^m w_i^{**} \tag{9.3b}$$

and

$$w_i^{**} = \left[\left(\frac{\hat{\bar{\rho}}_i}{\bar{r}_i} \right)^2 \left(\frac{S^2_{r_i}}{k_i} \right) \right]^{-1} \tag{9.3c}$$

and where $S^2_{r_i}$ is the variance of the observed correlations within a given meta-analysis, $\hat{\bar{\rho}}_i$ is the mean disattenuated correlation in that meta-analysis, $\bar{r}_i$ is the meta-analytic (bare-bones) mean correlation in that meta-analysis, $\hat{\bar{\bar{\rho}}}$ is the (weighted) grand mean effect size across the *m* meta-analyses, k_i is the number of primary studies included in the *i*th meta-analysis, and w_i^{**} is the weight applied to the *i*th meta-analysis. The

second term on the right side of Equation (9.3) is the weighted average second-order sampling error variance across the *m* meta-analyses:

$$E(S^2_{e_{\hat{\bar{\rho}}_i}}) = \sum_1^m \left[w_i^{**} \left(\frac{\hat{\bar{\rho}}_i}{\bar{r}_i} \right)^2 \frac{S^2_{r_i}}{k_i} \right] / \sum_1^m w_i^{**} \tag{9.3d}$$

Equation (9.3d) reduces to Equation (9.3e):

$$E(S^2_{e_{\hat{\bar{\rho}}_i}}) = m / \sum_1^m w_i^{**} \tag{9.3e}$$

The w_i^{**} are as defined in Equation (9.3c). Equation (9.3) has the same form as Equation (9.2), but some of the terms in it are estimated differently, so some explanation is indicated. The first term on the right side of Equation (9.3) is the computed variance across the meta-analyses of the first-order meta-analytic mean disattenuated population correlations. Computation of this value is shown in Equations (9.3a) and (9.3b). Equation (9.3c) shows the weights that are applied in Equations (9.3a) and (9.3b). The second term on the right in Equation (9.3) is the sampling error variance of these estimates. As shown in Equations (9.3d) and (9.3e), this sampling error is estimated as the weighted average across meta-analyses of the product of the square of the mean correction factor and the mean sampling error variance of the bare-bones (uncorrected) meta-analytic correlations ($S^2_{e_{\bar{r}}}$; see Equation [9.1d]). Each meta-analysis will have reported the variance of the observed correlations it included. Dividing this variance by *k* (the number of studies in the meta-analysis) yields the RE sampling error variance of the mean of the observed (uncorrected) correlations in that meta-analysis. As shown in Equations (9.3d) and (9.3e), the weighted average of the product of these values and the square of the correction factors across the *m* meta-analyses is the random effects sampling error variance estimate needed for Equation (9.3) (see Chapter 4). (As noted in the discussion of first-order artifact distribution-based meta-analysis in Chapter 4, this is based on the well-known principle that if one multiples a distribution of scores by a constant, the standard deviation is multiplied by that constant and the variance is multiplied by the square of that constant. Here the constant is the mean measurement error correction [$\hat{\bar{\rho}}_i / \bar{r}_i$].) The square root of the value on the left side of Equation (9.3d) divided by the square root of *m* is the standard error ($SE_{\hat{\bar{\bar{\rho}}}}$) and can be used to put confidence intervals around the grand mean ($\hat{\bar{\bar{\rho}}}$; computed in Equation [9.3b]).

The value on the left side of Equation (9.3) is the estimate of the nonartifactual variance of the population disattenuated correlations across the *m* meta-analyses. This is the variance remaining after subtraction of variance due to second-order sampling error. When this value is negative (i.e., second-order sampling error variance is greater than the observed variance across the first-order meta-analytic mean estimates), it is set to zero. Using the square root of this value ($\hat{\sigma}_{\bar{\rho}}$), credibility intervals around the grand mean correlation can be computed, as described earlier. If the value on the left side of Equation (9.3) is zero, the indicated conclusion is that these mean population correlations are the same across the *m* meta-analyses. All variance is accounted for by second-order sampling error, leading to the conclusion that there are no moderators. If this value is greater than zero, one can compute the proportion of between meta-analyses variance that is accounted for by second-order sampling error variance. This is computed as the ratio of the second term on the right side of Equation (9.3) to the first term; that is,

$$\text{ProportionVar} = E(S^2_{\hat{e}_{\hat{\rho}_i}}) / S^2_{\hat{\rho}} \tag{9.3f}$$

where 1 – ProportionVar denotes the proportion of the variance of the population disattenuated correlations that is true variance (i.e., variance not due to second-order sampling error). Because of this, this number is the reliability of the vector of mean corrected correlations across the *m* first-order meta-analyses. This follows from the fact that reliability is defined as the proportion of total variance that is true variance (i.e., variance not due to error; Magnusson, 1966; Nunnally & Bernstein, 1994). This reliability reflects the extent to which the mean first-order corrected correlations discriminate between the first-order meta-analysis results.

MIXED SECOND-ORDER META-ANALYSIS

In some cases, some of the first-order meta-analyses might have corrected each correlation individually while others applied the artifact distribution method. How, then, should the second-order meta-analysis be conducted? The meta-analyses that corrected each coefficient individually can be "converted" to artifact distribution meta-analyses, and the equations for second-order artifact distribution meta-analysis can be applied to all the first-order meta-analyses. The quantities needed in these equations (Equations [9.3] and [9.3a] through [9.3f]) are typically reported in meta-analyses that have corrected each correlation individually, making this conversion possible.

CONSIDERATIONS IN SECOND-ORDER META-ANALYSIS

One limitation of second-order meta-analysis methods is that the requirement for statistical independence of meta-analysis may limit the frequency with which the methods can be applied. The extent to which moderate violations of this assumption affect the results is unknown, but Cooper and Koenka (2012), in discussing an older, cruder form of second-order meta-analysis, suggest that minimizing the lack of independence might be sufficient to produce reasonably accurate results, and they give several examples of such published second-order meta-analyses. Tracz, Elmore, and Pohlmann (1992), in a simulation study, found that violations of the assumption of independence in first-order meta-analyses had minimal effect on the accuracy of results. Issues related to the importance of the independence assumption are discussed further in Chapter 10.

Second-order meta-analysis is not directly concerned with the variability of study population correlations *within* each of the first-order individual meta-analyses. To be sure, this variability within meta-analyses (i.e., nonartifactual variability between primary studies in first-order meta-analyses) is taken into account mathematically in second-order meta-analysis methods, as can be seen in Equations (9.1a), (9.1b), (9.1c), (9.2a), (9.2b), (9.2c), (9.3a), (9.3b), and (9.3c). But a finding that second-order sampling error accounts for all of the variability in the mean values across first-order meta-analyses does not imply that population parameters do not vary within first-order meta-analyses. Such a finding simply means that the *mean* values are equal across the different first-order meta-analyses. For example, the Schmidt and Oh (2013) finding that the mean meta-analytic operational validity for Conscientiousness is the same across different East Asian countries does not mean that this validity cannot vary somewhat across subpopulations within, for example, South Korea. If this is the case, this variability will be reflected in the results of the first-order meta-analysis. It is the purpose of the original first-order meta-analyses to address this nonartifactual variability between primary studies within each first-order meta-analytic context. The purpose of second-order meta-analysis is to gauge the true (nonartifactual) variability between meta-analyses (e.g., cross-country, cross-region, cross-criterion, cross-setting) for *mean values* of ostensibly the same relationship and to use this information to improve accuracy of estimation for each first-order meta-analytic mean estimate.

A possible objection to second-order meta-analysis is the following: Instead of second-order meta-analysis, why not conduct an overall meta-analysis, pooling all primary study data from all meta-analyses (which will yield the same grand mean as the second-order meta-analysis), and then break out into sub-meta-analyses based on hypothesized moderators (which yields the same subgroup means as those used in the second-order meta-analysis)? First, this is often an impossible or impractical alternative, because the primary studies used in all first-order

meta-analyses are often not available. Some journals (e.g., *Journal of Applied Psychology*) in the fields of organizational behavior and human resource management have only recently required that meta-analyses report all data from primary studies used in the meta-analysis (Aytug, Rothstein, Zhou, & Kern, 2012; Kepes, Banks, McDaniel, & Whetzel, 2012). As mentioned, second-order meta-analysis can be conducted using only first-order meta-analytic results (k, mean observed r, mean corrected r, and variance across observed or corrected rs), and thus it can be applied to most if not all previous first-order meta-analyses. Second, and perhaps more important, this procedure does not allow one to estimate the variance (and the percentage of variance) across subgroup meta-analyses that is (and is not) due to second-order sampling error variance, because second-order sampling error variance is not computed (or computable) in the omnibus meta-analysis approach. This is because omnibus meta-analyses and their subgroup meta-analyses are both first-order meta-analyses. For example, application of this approach to the Conscientiousness validity data in our first example would not have revealed that all the variance across the four East Asian countries in meta-analytic operational validity values was due to second-order sampling error. Instead, the values would have been taken at face value. So the omnibus meta-analysis procedure is not a substitute for second-order meta-analysis.

A variation on this objection is the following: Why not just conduct an omnibus, pooled meta-analysis along with subgroup meta-analyses based on hypothesized moderators and then look at the relative variances? The difference between the estimated population parameter variance in the omnibus meta-analysis and the average of this figure across the subgroup meta-analyses estimates the variance of the subgroup means (the variance of means across subgroup meta-analyses). This statement reflects the well-known ANOVA principle that total variance is the sum of between-group variance and average within-group variance. However, knowing the variance of the subgroup means does not allow one to estimate *how much* of this variance is (or is not) due to second-order sampling error and therefore does not allow computation of the proportion of this variance that is due to second-order sampling error. As a result, the analyses presented in the example in Schmidt and Oh (2013) cannot be conducted. For example, if all the between-mean variance was accounted for by second-order sampling error (as was the case with Conscientiousness in our first example application), there would be no way for one to know this. The procedure advocated here allows one to compute the percentage of *total variance* that is accounted for by between-group variance in mean values, but this is not the same as the percentage of between-group variance in mean values that is due to second-order sampling error variance. So again, this is a procedure that is not a substitute for second-order meta-analysis.

Another possible objection is this: Why not just compute a meta-regression in which coded hypothesized moderators are used to predict the primary study correlations pooled across all the first-order meta-analyses? (These correlations can be either observed correlations, as in bare-bones meta-analysis, or correlations corrected for measurement error.) This procedure fails for the same reason as above: The squared multiple correlation will reveal the percentage *of the total variance* that is accounted for by the hypothesized moderator or moderators. But it will not reveal the percentage of the variance in the mean values that is explained by second-order sampling error, and therefore the analyses allowed by second-order meta-analysis cannot be done. So this procedure is also not capable of being a substitute for second-order meta-analysis.

In conclusion, the methods of second-order meta-analysis provide unique information that cannot be obtained using the more traditional methods of first-order meta-analysis. The methods are particularly useful in conducting cross-culture generalization studies (i.e., synthesizing first-order meta-analyses conducted in different countries for the same relationship using within-country studies) and meta-analytic moderator analyses (i.e., comparing first-order meta-analytic results of the same relationship across different settings and/or groups; e.g., racial or social class groups). This unique information can be important from the point of view of cumulative knowledge and understanding, as illustrated in the several empirical examples presented in Schmidt and Oh (2013).

Second-Order Sampling Error: Technical Treatment

This section presents a more technical and analytical treatment of second-order sampling error and statistical power in meta-analysis. For the sake of simplifying the presentation, the results are presented for "bare-bones" meta-analyses, that is, meta-analyses for which sampling error is the only artifact that occurs and for which a correction is made. However, the principles apply to the more complete forms of meta-analysis presented in this book.

If a meta-analysis is based on a large number of studies, then there is little sampling error in the meta-analytic estimates. However, if the meta-analysis is based on only a small number of studies, there will be sampling error in the meta-analytic estimates of means and standard deviations. This is called second-order sampling error. There are potentially two kinds of second-order sampling error: sampling error due to incompletely averaged sampling error in the primary studies and sampling error produced by variation in effect sizes across studies. We will call unresolved sampling error from the primary studies "secondary second-order sampling error," or "secondary sampling error" for short. We will call sampling error due to variation in effect sizes "primary second-order sampling error." Table 9.1 shows the circumstances in

which the two types of second-order sampling errors occur. The key to this table is whether we have the homogeneous or heterogeneous case in the population. In the homogeneous case, there is no variance in ρ or δ in the population. In the heterogeneous case, the population values of ρ or δ do vary. As we noted in Chapters 5 and 8, the heterogeneous case is much more common in real data. Note that regardless of whether the set of studies is homogeneous or heterogeneous, there is always secondary second-order sampling error. This occurs because, in real data sets, it is never the case that the number of studies is infinite or that all studies have infinite sample size—the only conditions that can completely eliminate secondary second-order sampling error. However, primary second-order sampling error occurs only in the heterogeneous case. That is, when there is variance in ρ or δ, then primary second-order sampling error will be produced by the sampling of particular values of ρ_i or δ_i in individual studies. This cannot happen in the homogeneous case, because different values of ρ or δ cannot be sampled, because there is only a single value of ρ or δ in all studies. Because the homogeneous case is rare in real data, however, there will typically be both kinds of second-order sampling error in real meta-analyses. That is, typical real-world meta-analyses fall into the bottom row of Table 9.1.

Table 9.1 Second-order sampling error: Schematic showing when the two types of second-order sampling error occur.

	Secondary Second-Order Sampling Error	*Primary Second-Order Sampling Error*
Homogeneous Case		
$(S_\rho^2 = 0; S_\delta^2 = 0)$	Yes	No
Heterogeneous Case		
$(S_\rho^2 > 0; S_\delta^2 > 0)$	Yes	Yes

For simplicity, the following discussion will be written for analyses of the *d* statistic, but analyses based on correlations or other statistics are also subject to second-order sampling error when the number of studies is not large. In particular, second-order sampling error for correlations is directly analogous to that for *d* values.

Consider secondary sampling error. Meta-analytic estimates are averages. Thus, the sampling error in individual studies is averaged across studies. If enough studies are averaged, then the average sampling error

effects become exactly computable and, hence, exactly correctable. However, if the number of studies is small, then the average sampling error effects will still be partly random. For example, consider the mean effect size. Ignoring the small bias in the *d* statistic (see Chapters 7 and 8), the average *d* for the meta-analysis is

$$\text{Ave}(d) = \text{Ave}(\delta) + \text{Ave}(e) \tag{9.4}$$

If the number of studies is large, then the average sampling error across studies, Ave(*e*), will equal its population value of 0. That is, if we average across a large number of particular sampling errors, the sampling errors will cancel out exactly and yield an average of 0. If Ave(*e*) = 0, then

$$\text{Ave}(d) = \text{Ave}(\delta) \tag{9.5}$$

That is, if the average sampling error in the meta-analysis is 0, then the average observed effect size in the meta-analysis is equal to the average population effect size in the meta-analysis. If Ave(*e*) differs from 0, that is the effect of secondary sampling error.

If secondary sampling error were 0, then the average effect size in the meta-analysis would equal the average population effect size in the studies included in the meta-analysis. The number that we want to know, however, is the average population effect size across the entire research domain. The average effect size in the meta-analysis might differ from the average for the whole domain. If there were no variance in effect sizes across studies (the homogeneous case), then Ave(δ) = δ for any meta-analysis, and there can be no difference between the mean for the meta-analysis and the mean for the research domain. If there is variation across studies (the heterogeneous case), however, then the mean in the meta-analysis could differ by chance from the mean in the domain as a whole. This is primary second-order sampling error.

If the number of studies is large and if the studies are representative of the research domain, then the average population effect size in the meta-analysis, Ave(δ), will differ little from the average effect size across the research domain. That is, if the number of studies is large, then the Ave(*d*) value in the meta-analysis will be almost exactly equal to the average across the entire potential research domain. Thus, for a large number of studies, there will be no primary second-order sampling error in the meta-analysis mean.

In the next section, we will derive a confidence interval to estimate the potential range of second-order sampling error in the meta-analysis mean.

Both the mean (i.e., $\hat{\bar{\rho}}$ or $\hat{\bar{\delta}}$) and the standard deviation (i.e., SD_ρ or SD_δ) estimated in meta-analysis have second-order sampling error, although the exact relationship is more complicated in the case of standard deviations than it is for means. If the number of studies is large, then

the variance of the particular sampling errors in the meta-analysis, Var(e), will equal the value predicted from statistical theory. If the number of studies is small, then the observed sampling error variance may differ from the statistically expected value. Similarly, if the number of studies is large, then the variance in the particular effect sizes included in the meta-analysis, Var(δ), will equal the variance for the research domain as a whole. However, if the number of studies is small, then the variance of study population effect sizes in the meta-analysis may differ by chance from the variance of population effect sizes. This can also be stated as follows: If the number of studies is large, then the covariance between effect size and sampling error will be 0, but if the number of studies is small, then this covariance in the meta-analysis may differ by chance from 0.

Let us consider primary second-order sampling error in more detail. One key question is whether there is any primary second-order sampling error. There are two possible cases. First, there is the "homogeneous case" in which the population effect sizes do not differ from one study to the next (i.e., $S_\delta^2 = 0$). Second, there is the "heterogeneous case" where there is variation in population effect sizes across studies (i.e., $S_\delta^2 > 0$). Consider first the case in which the population study effect, δ_i, does not vary across studies. That is, in the homogeneous case, we have

$$\delta_i = \delta \text{ for each study } i \text{ in the domain}$$

As discussed in Chapters 5 and 8 and earlier in this chapter, the homogeneous case is probably rare in real data. In the homogeneous case, it is possible to speak of "the" population effect size δ. Because δ_i is the same for each study,

$$\begin{aligned} &\text{Ave}(\delta_i) = \delta \text{ for any set of studies from the domain} \\ &\text{Var}(\delta_i) = 0 \text{ for any set of studies from the domain} \end{aligned}$$

The meta-analysis mean observed effect size is

$$\begin{aligned} \text{Ave}(d_i) &= \text{Ave}(\delta_i) + \text{Ave}(e_i) \\ &= \delta + \text{Ave}(e_i) \end{aligned} \tag{9.6}$$

Thus, the meta-analytic average effect size differs from the effect size δ only to the extent that the average of the sampling errors in the meta-analysis differs from 0. That is, the only second-order sampling error in the mean effect size in the meta-analysis is the secondary sampling error, the sampling error resulting from primary sampling errors that by chance do not average to exactly 0.

In the homogeneous case, the population effect size is constant across studies. Thus,

$$\mathrm{Var}(d_i) = \mathrm{Var}(e_i)$$

If the number of studies were large, then the variance of the particular sampling errors in the meta-analysis would equal the variance predicted by the statistical theory for the research domain as a whole. However, if the particular sampling errors in the meta-analysis have a variance that is different by chance from the domain variance, then that unresolved primary sampling error will not have been eliminated from the meta-analysis. Thus, in the homogeneous case, the only second-order sampling error in the variance of observed effect sizes will be secondary sampling error, that is, unresolved first-order study sampling error.

Now let us consider the heterogeneous case in which population effect sizes *do* differ from one study to the next $(\text{i.e.}, S_\delta^2 > 0)$. The average observed effect size in a meta-analysis is

$$\mathrm{Ave}(d_i) = \mathrm{Ave}(\delta_i) + \mathrm{Ave}(e_i) \tag{9.7}$$

If the number of studies is small, then there can be error in each of the two terms: the average sampling error, Ave(e_i), and the average population effect size, Ave(δ_i). Consider the average sampling error, Ave(e_i). By chance, the average sampling error for that meta-analysis, Ave(e_i), is likely to depart from 0 by at least some small amount. That is secondary sampling error. Secondary sampling error always converges to 0 if the number of studies is large enough. However, it is possible for secondary sampling error to be small even if the number of studies is small. If the sample sizes in the primary studies were all very large—an unlikely event in psychological research—the average of the individual sampling errors would be near 0. The average sampling error would then be near 0 even though the number of studies is small.

Now consider the other term in the average effect size, Ave(δ_i), the average population effect size for the meta-analysis. If the number of studies is large, then the average population effect size in the meta-analysis will differ little from the average population effect size for the whole research domain. However, if the number of studies is small, then the particular values of (δ_i) observed in the meta-analysis are only a sample of the effect sizes from the domain as a whole. Thus, by chance, the average effect size in the meta-analysis may differ by some amount from the average effect size for the entire research domain. This departure is primary second-order sampling error. Even if all primary studies were done with an infinite number of subjects (i.e., even if every primary study sampling error e_i were 0), then the particular effect sizes in the meta-analysis need not have an average that is exactly equal to the domain average.

Thus, in the heterogeneous case, both the mean and the standard deviation of population effect sizes in the meta-analysis will depart from the research domain values because the studies observed are only a sample of studies. This is "primary second-order sampling error."

THE HOMOGENEOUS CASE

In defining the word *homogeneity*, it is important to distinguish between actual treatment effects and study population treatment effects. There are few studies that are methodologically perfect and, thus, few studies in which the study population treatment effect is equal to the actual treatment effect. In a research domain in which the actual treatment effect is the same for all studies, artifact variation across studies (e.g., varying levels of measurement error in different studies) will produce artifactual differences in study effect sizes. In most current textbooks on meta-analysis, the definition of *homogeneous* is obscured by implicit statistical assumptions. The definition of homogeneity requires that the study population effect sizes be exactly uniform across studies. In particular, most current chi-square homogeneity tests thus assume not only that the actual treatment effect is constant across studies but also that there is no variation in artifact values (e.g., measurement error) across studies. This assumption is very unlikely to hold in real data.

Most contemporary meta-analyses of experimental treatments have been bare-bones meta-analyses; no correction has been made for error of measurement or variation in strength of treatment, or variation in construct validity, or other artifacts. For a bare-bones meta-analysis, it is very unlikely that the study population effect sizes would be exactly equal for all studies. To have uniformity in the study effect sizes, the studies would have to be not only uniform in actual effect size but uniform in artifact values as well. All studies would have to measure the dependent variable with exactly the same reliability and the same construct validity. All studies would have to have the same degree of misidentification—inadvertent treatment failure—in group identification, and so on. (See the discussion of fixed vs. random meta-analysis models in Chapters 5 and 8; fixed effects meta-analysis models assume the homogeneous case; see also Chapters 2 and 6.) However, it may be useful in some cases to think of the homogeneous case as an approximation.

For purposes of this exposition of second-order sampling error, we assume homogeneity, and we denote the uniform study effect size by δ. Assume the average sample size to be 50 or more so that we can ignore bias in mean d values. Then, for each study individually, the treatment effect differs from δ only by sampling error. That is,

$$d_i = \delta + e_i \tag{9.8}$$

We then have

$$\text{Ave}(d) = \delta + \text{Ave}(e_i) \tag{9.9}$$

$$\text{Var}(d) = \text{Var}(e_i) \tag{9.10}$$

The average differs from δ only if the average sampling error is not the expected value of 0, that is, only if the number of studies is too low for errors to average out to the expected value (to within rounding error). The variance of observed effect sizes differs from Var(e) only if the variance of sampling errors Var(e_i) differs from the expected variance Var(e). This would not occur for a meta-analysis on a large number of studies. However, the sampling error in the variance estimate $(\hat{S}_\delta^2)$ is larger than the sampling error in the estimate of the mean $(\bar{\delta})$. Thus, in most meta-analyses, the sampling error in the estimate of the variance of effect sizes is much more important than the sampling error in the estimate of the average effect size.

In the homogeneous case, the sampling error in the mean effect size for a bare-bones meta-analysis is obtained from the sampling error equation

$$\bar{d} = \delta + \varepsilon$$

where $\bar{d}$ is the mean effect size and ε is the average sampling error. The distribution of meta-analytic sampling error ε is described by

$$E(\varepsilon) = 0$$

$$\text{Var}(\varepsilon) = \text{Var}(e) / K \tag{9.11}$$

where K is the number of studies and Var(e) is the average sampling error variance across the studies in the meta-analysis. $\text{Var}(\varepsilon) = SD_\varepsilon^2$. Thus, under the assumption of homogeneity, the 95% confidence interval for the mean effect size in a *bare-bones meta-analysis* is

$$\text{Ave}(d) - 1.96SD_\varepsilon < \delta < \text{Ave}(d) + 1.96SD_\varepsilon$$

(See Chapters 5 for and 8 methods of computing this confidence interval when artifacts beyond sampling error are corrected for.)

The sampling error in the estimated variance of effect sizes for a bare-bones meta-analysis is obtained by considering a variance ratio. For a large number of studies, the condition of homogeneity could be identified by computing the following ratio:

$$\text{Var}(d) \,/\, \text{Var}(e) = 1$$

For a small number of studies, this ratio will depart from 1 by sampling error. Many writers recommend that a chi-square test be used to assess the extent to which there is variance beyond sampling error variance. The statistic Q is defined as

$$Q = K\text{Var}(d) \,/\, \text{Var}(e)$$

We recommend that you not use the Q statistic. The Q statistic is the comparison variance ratio multiplied by the number of studies. Under the assumption of homogeneity, Q has a chi-square distribution with $K - 1$ degrees of freedom. This is the most commonly used "homogeneity test" of contemporary meta-analysis. The homogeneity test has all the serious flaws of any significance test. These flaws were discussed in Chapter 2. If the number of studies is small, then a real moderator variable must be enormous to be detected by this test. That is, the power of the test is low unless the moderator effect (interaction) is very large (Hedges & Pigott, 2001; National Research Council, 1992). On the other hand, if the number of studies is large, then any trivial departure from homogeneity, such as departures from artifact uniformity across studies, will suggest the presence of a moderator variable where there may be none. Because of these problems, we recommend against use of the homogeneity test.

THE HETEROGENEOUS CASE

If the research domain is heterogeneous $(\text{i.e., } S^2_\delta > 0)$, then there can be primary second-order sampling error—error due to the fact that the number of studies is not infinite. In a real meta-analysis in the heterogeneous case, there will therefore be two kinds of error: secondary sampling error and primary second-order sampling error. For purposes of discussion, we will focus first on just primary second-order sampling error. To do this, we will make a very unrealistic assumption: We will assume either (1) that all studies are done with infinite size or (2) (which is the same thing) that all study population effect sizes are known. After consideration of the special case, we will return to the realistic case of primary as well as second-order sampling error.

To make primary second-order sampling error clearly visible, let us eliminate first-order sampling error. That is, we assume all study Ns are infinite. Suppose population effect sizes do vary across studies (i.e., $S^2_\delta > 0$). The individual study effect size is δ_i. Under these assumptions, meta-analysis will compute the average and variance of the study effect sizes in the studies located:

$$\text{Ave}(d) = \text{Ave}(\delta_i)$$
$$\text{Var}(d) = \text{Var}(\delta_i)$$

However, if the number of studies is small, the average population effect size in the studies in the meta-analysis is only a sample average of the population effect sizes across all possible studies in the research domain.

The simplest case of a moderator variable is the binary case, for example, studies done with males versus studies done with females. The statistical description of a binary variable includes four pieces of information: the two values that are taken on by the binary variable and the probability of each value. Denote the two values by X_1 and X_2 and denote the respective probabilities by p and q. Because the sum of probabilities is 1, $p + q = 1$ and, hence, $q = 1 - p$. The mean value is

$$E(X) = pX_1 + qX_2 \tag{9.12}$$

Let D denote the difference between the values; that is, define D by

$$D = X_1 - X_2$$

The variance of the binary variable is

$$\text{Var}(X) = pqD^2 \tag{9.13}$$

Suppose a research domain has a moderator variable such that for 50% of studies, the effect size is $\delta = .20$, while for the other 50% of studies, the effect size is $\delta = .30$. For the research domain as a whole, the mean effect size is

$$\text{Ave}(\delta) = .50(.20) + .50(.30) = .25$$

The variance is given by

$$\text{Var}(\delta) = pqD^2 = (.50)(.50)(.30 - .20)^2 = .0025$$

Thus, the standard deviation is $SD_\delta = .05$. Consider a meta-analysis with $K = 10$ studies. If the studies are split 5 and 5, then for that meta-analysis, the mean effect size would be .25 and the standard deviation would be .05. Suppose, however, the studies by chance are split 7 and 3. The mean would be

$$\text{Ave}(d) = (7/10)(.20) + (3/10)(.30) = .23$$

rather than .25. The variance would be

$$\text{Var}(d) = (7/10)(3/10)(.30 - .20)^2 = (.21)(.01) = .0021$$

instead of .0025. That is, the standard deviation would be .046 rather than .05. These deviations in the mean and standard deviation of effect sizes are primary second-order sampling error, variation due to the fact that the sample of studies has chance variations from the research domain, which is the study population.

How large is primary second-order sampling error? The answer is simple for the mean effect size:

$$\text{Var}[\text{Ave}(\delta)] = \text{Var}(\delta) / K \tag{9.14}$$

The primary second-order sampling error variance of the variance estimate ($\hat{S}^2_\delta$) depends on the shape of the effect size distribution. That discussion is beyond the scope of the present book.

Consider now the case of a real meta-analysis with a small number of heterogeneous studies. There will be both primary and second-order sampling error. For the mean effect size in a bare-bones meta-analysis, each can be computed separately and easily:

$$\begin{aligned} \text{Var}[\text{Ave}(d)] &= \text{Var}[\text{Ave}(\delta)] + \text{Var}[\text{Ave}(e)] \\ &= \text{Var}(\delta) / K + \text{Var}(e) / K \\ &= [\text{Var}(\delta) + \text{Var}(e)] / K \\ &= \text{Var}(d) / K \end{aligned} \tag{9.15}$$

The square root of this quantity is the standard error of $\bar{d}$ and is used to create confidence intervals around $\bar{d}$. This formula holds for whatever set of weights is used in the basic estimation equations (see Hunter & Schmidt, 2000; Schmidt, Hunter, & Raju, 1988; Schmidt, Oh, & Hayes, 2009). Equation (9.15) applies to bare-bones meta-analysis; see Chapter 8 and 5 for methods of computing confidence intervals for $\bar{\hat{\delta}}$ or $\bar{\rho}$, respectively, in the heterogeneous case (random effects model) when measurement error and other artifacts in addition to sampling error are corrected for. The standard error of the standard deviation (or of $\hat{S}^2_\delta$) is much more complex and is beyond the scope of this book (cf. Raju & Drasgow, 2003).

A NUMERICAL EXAMPLE

Consider the first numerical example presented in Chapter 7:

N	*d*
100	.01
90	.41
50	.50
40	–.10

*Significant at the .05 level.

The meta-analysis using the more accurate formula found the following:

$$T = 280$$
$$K = 4$$
$$\bar{N} = 70$$
$$\text{Ave}(d) = .20$$
$$\text{Var}(d) = .058854$$
$$\text{Var}(e) = .059143$$

Here all observed variance is accounted for by sampling error, so the standard deviation of effect sizes is 0. Thus, the only second-order sampling error would be the secondary sampling error in the mean effect size. As described earlier, for the homogeneous case, the sampling error in the mean for a bare-bones meta-analysis is given by

$$\text{Var}[\text{Ave}(d)] = \text{Var}(e) / K = .059143 / 4 = .014786$$

and thus, the standard error of the mean is .12. The 95% confidence interval for the effect size δ is

$$.20 - 1.96(.12) < \delta < .20 + 1.96(.12)$$
$$-.04 < \delta < .44$$

Thus, the sampling error in this meta-analysis is substantial. We cannot be sure that the effect size is actually positive.

The problem in the previous meta-analysis is the total sample size. A total sample size of 280 would be a small sample size even for a single study. Thus, this meta-analysis can be expected to have considerable sampling error. To make this very explicit, suppose the number of studies was $K = 40$ rather than $K = 4$. The total sample size would then be $T =$ 2,800, which is far from infinite but still substantial. The sampling error variance would be

$$\text{Var}[\text{Ave}(d)] = \text{Var}(e) / K = .059143 / 40 = .001479$$

and the standard error would be .04. The confidence interval would be

$$.20 - 1.96(.04) < \delta < .20 + 1.96(.04)$$
$$.12 < \delta < .28$$

Thus, given 40 studies with an average sample size of 70, the average value of δ is known to be positive and the width of the 95% uncertainty interval shrinks from .48 to .16.

If the number of studies were 400, the total sample size would be 28,000 and the 95% confidence interval would shrink to

$$.18 < \delta < .22$$

Thus, under these assumptions, meta-analysis will eventually yield very accurate estimates of effect sizes. However, if the average sample size in the primary studies is very small, the number of studies required may be quite large.

ANOTHER EXAMPLE: LEADERSHIP TRAINING BY EXPERTS

Consider the leadership bare-bones meta-analysis from Table 7.1 in Chapter 7. Let us illustrate the computation of confidence intervals about those estimates. We have a heterogeneous case here, so we must use Equation (9.15) to compute the sampling error variance of our estimate of the mean. The sampling error variance in the mean effect size is

$$\text{Var}[\text{A}ve(d)] = \text{Var}(d) / K = .106000 / 5 = .021200$$

and the corresponding standard error is .146. The 95% confidence interval for the mean effect size is thus

$$.20 - 1.96(.146) < \text{Ave}(\delta) < .20 + 1.96(.146)$$
$$-.09 < \text{Ave}(\delta) < .49$$

This is a random effects standard error and a random effects confidence interval. Thus, with a total sample size of only 200, the confidence interval for the mean effect size is very wide.

This would also be true, however, for a single study with a sample size of only 200. For a single study with a sample size of 200 and an observed d of .20, the sampling error variance would be

$$\text{Var}(e) = [199 / 197][4 / 200][1 + .20^2 / 8] = .020304$$

The corresponding standard error would be .142, and the 95% confidence interval would be

$$.20 - 1.96(.142) < \delta < .20 + 1.96(.142)$$
$$-.08 < \delta < .48$$

The key to accuracy in the estimate of the mean effect size is to gather enough studies to generate a large total sample size.

For this example with a total sample size of 200, the 95% confidence interval for the mean effect size is $-.09 < \text{Ave}(\delta) < .49$. In particular, because the confidence interval extends below 0, we cannot be sure that the mean effect size is positive. On the other hand, it is equally likely to be off in the other direction. Just as the mean effect size might be .00 rather than the observed mean of .20, so with equal likelihood it could be .40 rather than the observed value of .20.

Assume now that we obtained similar results not for 5 studies but for 500 studies. For 500 studies with an average sample size of 40, the total sample size would be 500(40) = 20,000. There would be little sampling error in the meta-analysis estimates. The sampling error in the mean effect size would be

$$\text{Var}[\text{A}ve(d)] = \text{Var}(d)/K = .106000/500 = .000212$$

and the standard error would be .015. The 95% confidence interval for the mean effect size would be

$$.20 - 1.96(.015) < \text{Ave}(\delta) < .20 + 1.96(.015)$$
$$.17 < \text{Ave}(\delta) < .23$$

MODERATOR EXAMPLE: SKILLS TRAINING

Consider the overall bare-bones meta-analysis of the studies in Table 7.2 of Chapter 7. We have

$$T = 40 + 40 + \cdots = 400$$
$$K = 10$$
$$\bar{N} = T/10 = 40$$
$$\text{Ave}(d) = .30$$
$$\text{Var}(d) = .116000$$
$$\text{Var}(e) = [39/37][4/40][1 + .30^2/8] = .106591$$
$$\text{Var}(\delta) = .116000 - .106591 = .009409$$
$$SD_\delta = .097$$

This is again a heterogeneous case. The estimated standard deviation of effect sizes is .097, which is large relative to the mean of .30. However, the total sample size is only 400.

Because the total sample size is only 400, we should worry about the sampling error in the mean effect size. The sampling error in the mean effect size is thus

$$\text{Var}[\text{Ave}(d)] = \text{Var}(d)/K = .116000/10 = .011600$$

and the standard error is .108. The confidence interval for the mean effect size is thus

$$.30 - 1.96(.108) < \text{Ave}(\delta) < .30 + 1.96(.108)$$
$$.09 < \text{Ave}(\delta) < .51$$

That is, with a total sample size of 400, there is a large amount of sampling error in the mean effect size.

On the other hand, suppose we obtained these results not with 10 studies but with 1,000 studies. The total sample size would be 1,000(40) = 40,000, and there would be very little sampling error in the mean effect size. The 95% confidence interval for the mean effect size would be

$$.30 - 1.96(.0108) < \text{Ave}(\delta) < .30 + 1.96(.0108)$$
$$.28 < \text{Ave}(\delta) < .32$$

Confidence Intervals in Random Effects Models: Hunter-Schmidt and Hedges-Olkin

The way in which the standard error of the mean *r* or *d* is estimated in a random effects meta-analysis differs between the methods presented in this book and the method presented by Hedges and Vevea (1998).

Estimation procedures are simpler for the Hunter-Schmidt (H-S) approach (Schmidt, Hunter, & Raju, 1988), so we present those procedures first.

The Hunter-Schmidt Random Effects (RE) Procedure. Our presentation is in terms of the *d* statistic, but procedures are similar and analogous for *r* and other indices of effect size. In the H-S RE procedure, the sampling error variance of the mean *d* is estimated as the variance of the observed *d*s across studies divided by *k*, the number of studies:

$$S^2_{e_{\bar{d}}} = \frac{\bar{V}_e}{k} + \frac{S^2_\delta}{k} = \frac{S^2_d}{k} \tag{9.16}$$

The square root of Equation (9.16) is the *SE* that is used in computing CIs:

$$SE_{\bar{d}} = \frac{SD_d}{\sqrt{k}} = \sqrt{\frac{\bar{V}_e + S^2_\delta}{k}} \tag{9.17}$$

In this model, $\bar{V}_e$ is conceptualized as the sample size weighted mean of the V_{e_i} values. The equation for S^2_d is

$$S^2_d = \sum N_i (d_i - \bar{d})^2 \ / \ \Sigma N_i \tag{9.18}$$

where

$$\bar{d} = \Sigma N_i d_i \ / \ \Sigma N_i \tag{9.19}$$

The rationale for this procedure can be seen in the fact that $S^2_d = S^2_e + S^2_\delta$. That is, the expected value of S^2_d is the sum of simple sampling error variance and the variance of the study population parameters (Chapters 3 and 7; Field, 2005; Hedges, 1989). Hence, S^2_d divided by *k* is the sampling

error variance of the mean. Osburn and Callender (1992) showed that this equation holds both when $S_\delta^2 > 0$ and when $S_\delta^2 = 0$ (i.e., when the assumption underlying the FE model holds). The study weights in the H-S RE model are (total) study sample sizes, N_i, used because these weights closely approximate the inverse of the simple sampling error variances ($1 / V_{e_i}$) (see Chapter 3) and are less affected by sampling error variance (Brannick, 2006). Hedges (1983b) stated that in the heterogeneous case ($S_\delta^2 > 0$), weighting by sample size "will give a simple unbiased estimator [of the mean] that is slightly less efficient than the optimal weighted estimator" (p. 392). Osburn and Callender (1992) showed via simulation that weighting by sample size produces accurate *SE* estimates both when $S_\delta^2 = 0$ and when $S_\delta^2 > 0$. Also using simulation, Schulze (2004) found that for heterogeneous population data sets, the H-S RE procedure weighting by sample size produced accurate (more accurate than other procedures evaluated) estimates of CIs (see his Table 8.13, p. 156); estimates for the mean correlation were also acceptably accurate (with a tiny median negative bias of .0022, much less than rounding error; Table 8.4, p. 134; see pp. 188–190 for a summary). Brannick (2006) reported similar results. Further details can be found in Osburn and Callender (1992) and Schmidt, Hunter, and Raju (1988). We note here that in the H-S RE method, when the *d*s are corrected for measurement error, the procedure is analogous except that S_d^2 is now the variance of the corrected *d*s. The same is true for *r* value meta-analyses. Standard errors of the mean for corrected mean values are given in Chapter 5 for *r* values and Chapter 8 for *d* values. The Hedges-Vevea (H-V) procedure does not include corrections for artifacts.

The Hedges-Vevea RE Procedure. The Hedges and Vevea (1998) RE procedure estimates the two components of RE sampling error variance separately. The simple sampling error variance component is estimated exactly as it is in the FE model:

$$S_{e_{\bar{d}}}^2 = 1 / \Sigma w_i \tag{9.20}$$

where the w_i are $1 / V_{e_i}$.

The second component, $\hat{\sigma}_\delta^2$ (symbolized as $\hat{\tau}^2$ by Hedges and Vevea), is estimated as follows:

$$\hat{\sigma}_\delta^2 = \begin{cases} \dfrac{Q-(k-1)}{c} & \text{if } Q \geq k-1 \\ 0 & \text{if } Q < k-1 \end{cases} \tag{9.21}$$

where $Q = \chi^2$ overall homogeneity test and c is a function of the study weights and is given in Equation (11) from Hedges and Vevea (1998):

$$c = \sum w_i - \frac{\sum (w_i)^2}{\sum w_i} \tag{9.22}$$

where the study weights w_i are the FE study weights as defined in our Equation (9.20).

The estimated mean value is then

$$\hat{\delta} = \bar{d} = \Sigma w_i^* d_i \ / \ \Sigma w_i^* \tag{9.23}$$

The sampling error variance is

$$S^2_{e_{\bar{d}}} = 1 \ / \ \Sigma w_i^* \tag{9.24}$$

where the w_i^* are $1 \ / \ [V_{e_i} + \hat{\sigma}^2_{\delta}]$.

When the effect size statistic is the correlation, this RE procedure first converts *r*s to the Fisher's *z* transformation, conducts the calculations in that metric, and then back transforms the resulting means and CIs into the *r* metric (Hedges & Olkin, 1985). The Fisher's *z* transform is discussed in Chapter 5. See Hedges and Vevea (1998), Field (2005), and S. M. Hall and Brannick (2002) for a complete technical description of this RE procedure.

$\hat{\sigma}^2_{\delta}$ in Equation (9.21) is set to zero when $Q - (k - 1)$ yields a negative value, because by definition a variance cannot be negative. Hedges and Vevea (1998) discuss the positive bias that characterizes this estimate as a result of setting negative values to zero, and they tabulate this bias in their Table 2 for various conditions. This bias causes the *SE* to be upwardly biased, causing the resulting CIs to be too wide; that is, the probability content of the CIs is larger than the nominal value (Hedges & Vevea, 1998, p. 496). Overton (1998, pp. 371, 374) found this same bias for this procedure and also for an iterative procedure he used to estimate S^2_{ρ} and S^2_{δ}. Hedges and Vevea state that bias becomes smaller as *k* (the number of studies) increases and is generally small when *k* is 20 or more. However, Overton (1998) pointed out that the bias also depends on the actual size of S^2_{δ} (or S^2_{ρ}). For example, if this value is zero, then 50% of the estimates are expected to be negative due to sampling error, creating a positive bias regardless of the number of studies. If this value is small but not zero, then less that 50% of the estimates of S^2_{δ} are expected to be negative, and the positive bias is smaller. When S^2_{δ} is large, the positive bias is negligible. Overton (1998) stated that when S^2_{δ} is small, the RE model overestimates sampling error variance and produces CIs that are too wide. This effect is not due to any inherent property of the RE model; it is due to the positive bias in the procedures he examined for estimating the standard error of the mean

meta-analysis value. Some researchers have mistakenly cited Overton's statement as a rationale for preferring the FE model to the RE model in their meta-analyses (e.g., Bettencourt, Talley, Benjamin, & Valentine, 2006).

Because of its different mode of estimating the sampling error variance (described earlier), the H-S RE procedure does not have this upward bias. As shown earlier, in the H-S RE procedure, the two components of the RE sampling error variance are estimated jointly rather than separately. Note that if S_δ^2 is in fact zero, the H-S RE estimate of sampling error variance has the same expected value as the FE estimate of sampling error variance (Osburn & Callender, 1992; Schmidt, Hunter, & Raju, 1988; Schmidt, Oh, & Hayes, 2009). As shown by Hedges and Vevea (1998), this is not the case for the H-V RE procedure.

Updating a Meta-Analysis When a New Study Becomes Available

When a new study becomes available, there are two ways in which one can update the meta-analysis to include this study. First, one can rerun the meta-analysis including the new study. Second, one can take a Bayesian approach. In that approach, one would treat the existing fully corrected meta-analysis mean and *SD* as the Bayesian prior distribution and multiply this distribution times the likelihood function from the new study, using the usual Bayesian equation. The likelihood function or distribution has as its mean the fully corrected *r* or *d* value from the new study, and its *SD* is the standard error (*SE*) of that estimated corrected *r* or *d* value (i.e., the square root of the sampling error variance of the corrected value from the study). Either of these procedures can also be applied when there are multiple new studies. Schmidt and Raju (2007) have examined the properties of these two procedures in detail. They conclude that it is virtually always best to rerun the meta-analysis including the new study or studies.

What Are Optimal Study Weights in Random Effects Meta-Analysis?

Considerable attention has been devoted in the literature to the question of how the studies in a meta-analysis should be weighted. In the H-V procedure, because of the nature of the study weights used to produce the weighted mean *d* value (or *r* value), it is necessary when using these weights to have a separate estimate of σ_δ^2 (Field, 2005; Hedges & Vevea, 1998). As noted earlier, the weight applied to each study is $w_i^* = 1 / [V_{e_i} + \hat{\sigma}_\delta^2]$, where V_{e_i} is the simple sampling error variance for that study. The H-S procedure weights each study by its (total) sample size (N_i) and

therefore does not require a separate estimate of σ^2_δ. (As noted in Chapter 3, when correlations are corrected individually, the H-S procedure weights studies by the product of the study *N* and the compound attenuation factor.) Of course, the H-S RE model does estimate σ^2_δ for other purposes (such as credibility intervals), and this estimate does have a positive bias (discussed in Chapter 5), but this estimate is not used in the weights applied to the studies and so does not affect the computation of weighted mean values, *SE*s, or confidence intervals (Schmidt, Hunter, & Raju, 1988; Schmidt, Oh, & Hayes, 2009; see also Schulze, 2004, p. 190). The H-V weights were derived within the context of large sample statistical theory, that is, under the assumption the number of studies and study *N*s are very large. In such a hypothetical situation, the H-V study weights are in expectation more accurate for RE models (Hedges, 1983a, 1983b; Hedges & Vevea, 1998; Raudenbush, 1994; Schulze, 2004, 2007). But even within large sample theory, this advantage is slight (Hedges, 1983b, p. 393). The problem in using these weights with actual data is that the small theoretically expected advantage for these study weights is not realized with the smaller numbers of studies and study *N*s that are typical in real meta-analyses, because of inaccuracies induced by sampling error in the estimates of the σ^2_δ component of the weights (e.g., see Brannick, 2006; Raudenbush, 1994, p. 317; and Schulze, 2004, pp. 84 and 184; 2007). Because of this effect, Schulze (2004, pp. 193–194), based on the results of his extensive Monte Carlo studies, recommended weighting studies by sample size in the heterogeneous case (i.e., σ^2_δ or $\sigma^2_\rho > 0$), as well as the homogeneous case. Kulinskaya, Morgenthaler, and Staudte (2010) reached this same conclusion, as did Shuster (2009). Brannick (2006) conducted an extensive simulation study in the *r* metric. He found that sample size study weights produced estimates that were less biased and had smaller root mean square error than weighting by inverse sampling error variances. He concluded that the accuracy problems of the inverse study weights stemmed from the fact that sampling error often causes the *r* statistic to take on extreme values, which cause extreme study weights, which, in turn, cause inaccurate estimates of mean correlations. In a later study, Brannick, Yang, and Cafri (2011) confirmed these results favoring *N*-weighting for the *r* metric but found that in the case of the *d* metric, weighting by inverse variance had a slight advantage over weighting by *N*. Marin-Martinez and Sanchez-Meca (2010) also reported this result. However, Sanchez-Meca and Marin-Martinez (1998), in another simulation study in the *d* metric, found that weighting studies by sample size resulted in unbiased estimates of mean *d* under all conditions, while inverse sampling variance weights produced slightly (negatively) biased estimates. At the same time, weighting by inverse variance was slightly (2.8%) more statistically efficient. In the case of the *d* metric, the differences between the two

weighting methods appear to be very tiny and not of practical significance in research.

The random effects study weights used in the Hedges procedure are less unequal across studies than sample size weights. Hence, that procedure gives relatively more weight to studies based on small sample sizes, which can cause problems in applying widely used methods for the detection of publications bias (see Chapter 13).

Bonett (2008, 2009) has challenged both these approaches to weighting of studies when used with the RE meta-analysis model. He argues that the RE model is based on the assumption that the studies in the meta-analysis are a random sample of a defined population of studies and that this assumption cannot be justified because meta-analysts cannot appropriately define or delimit such a population. Because of this problem, he advocates that all studies be weighted equally. He is correct that in the RE model, the studies in the meta-analysis are viewed as a random sample from a larger universe of studies that exist or could be conducted. Hedges and Vevea (1998) pointed out that this larger universe is often poorly defined and ambiguous in nature. However, Schulze (2004, pp. 40–41) noted that this is not a problem specific to meta-analysis or RE models in meta-analysis but one that characterizes virtually all samples used in primary and other research. Rarely in research is the target population of subjects fully enumerated and delimited; in fact, data sets used frequently consist of something close to convenience samples (i.e., a set of subjects for whom it was possible to obtain data). Viewed in this light, this problem appears less serious. We can ask how different meta-analytic results would be using equal study weights. Brannick et al. (2011) evaluated equal weights in their simulation study. They found that both sample size weights and inverse variance weights were more accurate and efficient than equal weights, but the differences were often negligible from a practical point of view. However, this study does not speak directly to Bonett's (2008, 2009) objection, because in the Brannick et al. simulation study, there was in fact a clearly defined population of studies from which studies were sampled.

The Meaning of Percent Variance Accounted for in Meta-Analysis

In the first part of Chapter 5, we presented the case against percent variance accounted for as a useful statistic in any kind of research. Yet in Chapters 3, 4, and 7, we often presented figures for the percentage of variance in *r* or *d* values accounted for by sampling error and other artifacts in meta-analysis. In doing this, we have tried to point out a more meaningful interpretation of this meta-analytic result: The square root of the

proportion of variance explained is the correlation between the observed *r* or *d* values, on one hand, and the sampling errors and other artifactual perturbations in the effect sizes, on the other. For example, if 81% of the variability across effect sizes is explained by artifacts, then the observed effect sizes are correlated .90 (the square root of .81) with artifact-produced perturbations in the observed values. If 50% is accounted for, this correlation is .71 ($r = \sqrt{.50}$). The correlation is much easier for readers and research users to understand than percent variance. Correlations (i.e., linear relations) exist in the real world while variances do not; a variance is a quadratic statistic created by squaring data points of interest and is in that sense artificial. In addition, as noted in Chapter 5, the percent variance statistic is highly susceptible to misinterpretation, because small percent variance figures are often wrongly dismissed as unimportant when the effect sizes underlying them are fairly large and of practical significance. So we recommend that in meta-analysis, the final percent variance accounted for figures be converted to correlations.

The Hedges and associates meta-analysis methods as presented in Borenstein et al. (2009) include an index of percent variance called I^2, which is attributed to Higgins et al. (2003). This index represents the proportion of variance *not* explained by sampling error (the only artifact addressed by that method). In a Hunter-Schmidt bare-bones meta-analysis, 1 minus the proportion of variance accounted for equals I^2. Both indices are affected by the *N*s of the studies in the meta-analysis, because sampling error causes most artifactual variance. Other things constant, if study *N*s are small, percent variance explained tends to be large, I^2 tends to be small, and the correlation between observed values and perturbations due to artifacts tends to be large. The opposite tends to be the case when the *N*s of the studies in the meta-analysis are large. This dependence on study *N*s should be borne in mind in interpreting both these indices.

It is also important to remember that the proportion of variance explained is less informative when the observed variance of the meta-analytic correlations or *d* values is small. A percent-based estimate can be misleading when it is interpreted blindly without considering the size of its denominator. For example, a proportion of variance figure of 50% could be .1000 / .2000 or .00010 / .00020. The latter case would not suggest the existence of moderator(s), given the tiny amount of observed variation to begin with and the even smaller amount of nonartifactual variance. For purposes of detecting the likely presence of moderators, the absolute amount of true variance (nonartifactual variance) in the study effects sizes (or even better, its square root, the *SD*) can be more important than the relative percent of variance attributable to artifacts. We suggest that meta-analysts consider both estimates.

The Odds Ratio (OR) in Behavioral Meta-Analyses

In medical research, both the independent and dependent variables are often true dichotomies—for example, vaccinated versus not vaccinated (independent variable) and contracted the disease versus did not contract the disease (dependent variable), creating a 2 × 2 table. The favored and most widely used ratio measure in medical research is the odds ratio (Haddock et al., 1998). The odds ratio (OR) is the ratio of two probabilities:

$$OR = P(I/E)/ P(I'E),$$

where P(I/E) (in our example) is the probability of getting the disease in the group that did not get the vaccine, and P(I'E) is the probability of getting the disease in the group that got the vaccine. These probabilities are estimated via ratios between cells in the 2 × 2 table. Primary studies and meta-analyses using the OR statistic analyze the natural logs of the ORs (ln(OR)), with the final results then being converted back to the OR metric. The OR is seldom used in psychological or behavioral research because it is rare that both variables are true dichotomies. In fact, in many studies, both variables are continuous; this is especially frequent in correlational studies. Of course, in experiments, the independent variable is often dichotomous: the treatment group versus the control group. However, the dependent variable is almost always continuous or at least not dichotomous. Some examples are amount learned in the training program, degree of change in racial attitudes, and amount of reduction in anxiety. It is rare in behavioral research to have a truly dichotomous dependent variable. Of course, dependent variables that are actually continuous can be artificially dichotomized to allow application of the OR, but it is well known that doing this is not good practice because it causes a major loss of information (Cohen, 1983; Hunter & Schmidt, 1990a; MacCallum et al., 2002). In their explication of the OR, Haddock et al. (1998) present an example in which the independent variable is a psychosocial treatment for drug addiction (experimental vs. control group) and the dependent variable is "successful" or "not successful" in reducing drug use. This is an example of what MacCallum et al. (2002) warned against: an artificial dichotomization of a continuous variable. There are degrees of success in reducing drug usage. We believe that this sort of consideration is the reason why the use of the OR is still rare in behavioral research 15 years after Haddock et al. (1998) advocated its use for behavioral research in the journal *Psychological Methods*.

There is another important reason to avoid using the OR: Most people find it difficult to understand the meaning of an OR—not only laypeople but also medical practitioners (to whom the medical research is directed)

and even medical researchers themselves. The OR is not an intuitive statistic (Borenstein et al., 2009). In fact, as demonstrated by Gigerenzer and his associates (Gigerenzer, 2007; Gigerenzer et al. 2007), the vast majority of practicing medical doctors routinely seriously misinterpret the meaning of outcome statistics used in medical research. Patients would be quite concerned if they were aware of this fact.

Suppose some of the studies relevant to one's meta-analysis present results in the form of ORs. Must these studies be omitted? Actually, it is quite easy to covert OR to either *d* or *r* values. These conversion formulas are given by Bonett (2007), Borenstein et al. (2009), and Chinn (2000). Both OR values and their sampling error variances can be converted. So such studies can be included in a meta-analysis conducted in the *r* or *d* metrics. If *all* relevant primary studies use the OR statistic, all can be converted to the *r* or *d* metric prior to meta-analysis. This is the procedure we recommend in both cases.

Exercise 9.1: Second-Order Meta-Analysis Across Different Independent Variables With the Same Dependent Variable

This exercise is based on the data used in the exercise at the end of Chapter 4. That exercise required you to conduct separate bare-bones meta-analyses for each of six tests. This results in an estimate of the percentage of variance accounted for by sampling error for each test. These are the first set of figures needed for this exercise. We hope you have retained them.

These data meet the requirements for a second-order meta-analysis, as described in this chapter. That is, the six meta-analyses are very similar substantively, and there is no reason to believe that there are different nonartifactual sources of variance (i.e., moderators) for the different tests.

Conduct a second-order meta-analysis across these six tests using the methods described in this chapter for second-order meta-analysis across different in dependent variables.

What is the average percentage of variance accounted for by sampling error across these six tests? What is your interpretation of this finding?

In the exercise at the end of Chapter 4, you also computed the percentage of variance accounted for by *all* artifacts—sampling error plus the other artifacts. This was computed not as part of the bare-bones meta-analysis but as part of the full meta-analyses, which corrected for measurement error and range restriction, as well as for sampling error. There were two such meta-analyses—one correcting for direct and one correcting for indirect range restriction. Conduct a separate second-order meta-analysis for each of these sets of percentage of variance figures.

Are these values different from those computed earlier based only on sampling error variance? Why is this difference in this particular set of data not larger than it is? What is your interpretation of these values?

Exercise 9.2: Second-Order Meta-Analysis With Constant Independent and Dependent Variables

Chanchal Tamrakar (2012) conducted separate, independent meta-analyses on the relationship between customer satisfaction and customer loyalty for (1) the retail industry, (2) the tourism industry, and (3) the telecom industry. He also conducted separate, independent meta-analysis by geographical area: (1) Asia, (2) Europe, and (3) North America. He used the artifact distribution method of meta-analysis in all six of his meta-analyses. His first-order meta-analysis results that you need for this exercise are shown in the first four columns in the following table. Conduct two second-order meta-analyses on his results—one for the industry categories and one for the geographical areas. To conduct these analyses, you need to use the following equations: (9.3) and (9.3a) through (9.3f) (a total of seven equations). Columns 5 through 12 are for your answers. The headings on these columns are as defined in the discussion in the text of equations listed here. They are also defined in the notes to the table.

Explain your results.

1. How much different are the mean corrected correlations after adjustment for second-order sampling error (column 12) in comparison with the values originally reported by Tamrakar (column 4)? Make this comparison separately for the two sets of meta-analyses.
2. Compare the percent of variance in his mean corrected correlations explained by second-order sampling error (column 10) for the industry category versus the geographical region category. What might be the explanation for this difference?

Exercise 9.2 Second Order Meta-Analysis of the Relationship between Customer Satisfaction and Customer Loyalty (Tamrakar, 2012, Table 1)

Predictor	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Moderator	k	$\bar{r}_i$	S_r^2	$\hat{\bar{\rho}}_i$	$S^2_{e_{\hat{\bar{\rho}}_i}}$	$\hat{\bar{\bar{\rho}}}$	$E(S^2_{e_{\hat{\bar{\rho}}_i}})$	$S^2_{\hat{\bar{\rho}}}$	$\sigma^2_{\hat{\bar{\rho}}}$	ProVar	$r_{\rho\rho}$	$\hat{\bar{\rho}}_{ir}$
Industry												
Retail	10	.55	.01605	.67								
Tourism	14	.69	.01415	.85								
Telecom	7	.59	.00335	.71								
Region												
Asia	13	.59	.02503	.72								
Europe	12	.69	.00881	.83								
North America	15	.61	.02211	.72								

Note. Columns (1) though (4) are input values (italicized) available from first order meta-analyses. (1) Number of samples; (2) Sample size weighted mean observed validity; (3) Sample size weighted observed variance across observed validities; (4) First order meta-analytic mean validity estimates; ; (5) Second order sampling error variance for each first order meta-analytic validity estimate (see discussion of Eq. 7d); (6) Second order, grand mean validity estimates (Eq. 7b); (7) Expected (average) second order sampling error variance (Eq. 7d) and standard error (in parentheses); (8) Observed variance and SD (in parentheses) across first order mean operational validity estimates (Eqs. 7a, 7b, and 7c); (9) Estimated true variance and SD (in parentheses) across first order mean operational validity estimates after expected second order sampling error variance is subtracted out from the observed variance (Eq. 7); negative values are set to zero; (10) The proportion (percentage if multiplied by 100) of the observed variance across first order mean operational validity estimates that is due to second order sampling error variance; values greater than 1 are set to 1; (11) The reliability of the first order meta-analytic validity vectors; these values are computed as 1 minus the values in Column 10; (12) Regressed first order validity estimates based on the reliability of the original validity vectors shown in Column 11.

Exercise 9.3: Second-Order Meta-Analysis With Constant Independent and Dependent Variables

Van Iddekinge, Roth, Putka, and Lanivich (2011) meta-analyzed relationships between job-related interests and job turnover. They examined this relationship for three different types of interest measures: (1) job and vocation focused scales, (2) construct-focused interest scales, and (3) basic interest scales. They also meta-analyzed the relationship between job interests and three different types of turnover: (1) voluntary, (2) involuntary, and (3) "other turnover." In all six of these first-order meta-analyses, they used the artifact distribution meta-analysis method. Their first-order meta-analysis results that you need for this exercise are shown in the first four columns in the following table. Conduct two second-order meta-analyses on their results—one for the type of interest scale meta-analyses and one for type of turnover meta-analysis. To conduct these analyses, you need to use the following equations: (9.3) and (9.3a) through (9.3f) (a total of seven equations). Columns 5 through 12 are for your answers. The headings on these columns are as defined in the discussion in the text of equations listed here. They are also defined in notes to the table.

Explain your results.

1. How much different are the mean corrected correlations after adjustment for second-order sampling error (column 12) in comparison with the values originally reported by Iddekinge et al. (2011) (column 4)? Compare separately for the two sets of first-order meta-analyses.
2. Compare the percent of variance in the Iddekinge et al. mean corrected correlations explained by second-order sampling error (column 10) for the type of interest scale category versus the type of turnover category. Why do you think these values are so different?

Exercise 9.3 Second Order Meta-Analysis of the Relationship between Customer Satisfaction and Customer Loyalty (Tamrakar, 2012, Table 1)

Predictor	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Moderator	k	$\bar{r}_i$	S_r^2	$\hat{\bar{\rho}}_i$	$S^2_{e_{\hat{\bar{\rho}}_i}}$	$\hat{\bar{\bar{\rho}}}$	$E(S^2_{e_{\hat{\bar{\rho}}_i}})$	$S^2_{\hat{\bar{\rho}}}$	$\sigma^2_{\bar{\rho}}$	ProVar	$r_{\rho\rho}$	$\hat{\bar{\rho}}_{ir}$
Type of Interest scale												
Job and vocation focused	10	–.16	.00467	–.17								
Construct focused	11	–.11	.00602	–.12								
Basic Interest scales	6	–.11	.00280	–.13								
Nature of Turnover												
Voluntary	15	–.20	.01363	–.22								
Involuntary	2	–.13	.00004	–.15								
Other turnover	15	–.11	.00351	–.11								

Note. CFP = corporate financial performance; CSP = corporate social/environmental performance. Columns (1) though (4) are input values (italicized) available from first order meta-analyses. (1) Number of samples; (2) Sample size weighted mean observed validity; (3) Sample size weighted observed variance across observed validities; (4) First order meta-analytic mean validity estimates; (5) Second order sampling error variance for each first order meta-analytic validity estimate (see discussion of Eq. 7d); (6) Second order, grand mean validity estimates (Eq. 7b); (7) Expected (average) second order sampling error variance (Eq. 7d) and standard error (in parentheses); (8) Observed variance and SD (in parentheses) across first order mean operational validity estimates (Eqs. 7a, 7b, and 7c); (9) Estimated true variance and SD (in parentheses) across first order mean operational validity estimates after expected second order sampling error variance is subtracted out from the observed variance (Eq. 7); negative values are set to zero; (10) The proportion (percentage if multiplied by 100) of the observed variance across first order mean operational validity estimates that is due to second order sampling error variance; values greater than 1 are set to 1; (11) The reliability of the first order meta-analytic validity vectors; these values are computed as 1 minus the values in Column 10; (12) Regressed first order validity estimates based on the reliability of the original validity vectors shown in Column 11.