Meta-Analysis of Correlations Corrected Individually for Artifacts

Introduction and Overview

In Chapter 2, we examined 11 study design artifacts that can affect the size of the correlation coefficient. At the level of meta-analysis, it is possible to correct for all but one of those artifacts: reporting or transcriptional error. Except for outlier analysis, we know of no way to correct for data errors. Outlier analysis can detect some but not all bad data and is often problematic in meta-analysis (as discussed in Chapter 5). Sampling error can be corrected for, but the accuracy of the correction depends on the total sample size that the meta-analysis is based on. The correction for sampling error becomes perfect as total sample size approaches infinity. Our discussion of meta-analysis in this chapter and in Chapter 4 will implicitly assume that the meta-analysis is based on a large number of studies. If the number of studies is small, then the formulas presented here still apply, but there will be nontrivial sampling error in the final meta-analysis results. This is the problem of “second order” sampling error, which is discussed in Chapter 9.

The 10 potentially correctable study design artifacts are listed in Table 3.1. To correct for the effect of an artifact, we must have information about the size and nature of the artifact. Ideally, this information would be given for each study (i.e., each correlation) individually for each artifact. In that case, each correlation can be corrected individually, and the meta-analysis can be conducted on the corrected correlations. This type of meta-analysis is the subject of this chapter.

Artifact information is often available only on a sporadic basis and is sometimes not available at all. However, the nature of artifacts is such that, in most research domains, the artifact values will be independent across studies. For example, there is no reason to suppose that reliability of measurement will be either higher or lower if the sample size is large or small. If the artifacts are independent of each other and independent of the size of the true population correlation, then it is possible to base meta-analysis on artifact distributions. That is, given the independence
Table 3.1  Study artifacts that alter the value of outcome measures
(with examples from personnel selection research)

1. Sampling error:
   Study validity will vary randomly from the population value because of sampling error.

2. Error of measurement in the dependent variable:
   Study validity will be systematically lower than true validity to the extent that job performance is measured with random error.

3. Error of measurement in the independent variable:
   Study validity for a test will systematically understate the validity of the ability measured because the test is not perfectly reliable.

4. Dichotomization of a continuous dependent variable:
   Turnover, the length of time that a worker stays with the organization, is often dichotomized into “more than...” or “less than...” where ... is some arbitrarily chosen interval such as one year or six months.

5. Dichotomization of a continuous independent variable:
   Interviewers are often told to dichotomize their perceptions into “acceptable” versus “reject.”

6. Range variation in the independent variable:
   Study validity will be systematically lower than true validity to the extent that hiring policy causes incumbents to have a lower variation in the predictor than is true of applicants.

7. Attrition artifacts: Range variation in the dependent variable:
   Study validity will be systematically lower than true validity to the extent that there is systematic attrition in workers on performance, as when good workers are promoted out of the population or when poor workers are fired for poor performance.

8. Deviation from perfect construct validity in the independent variable:
   Study validity will vary if the factor structure of the test differs from the usual structure of tests for the same trait.

9. Deviation from perfect construct validity in the dependent variable:
   Study validity will differ from true validity if the criterion is deficient or contaminated.

10. Reporting or transcriptional error:
    Reported study validities differ from actual study validities due to a variety of reporting problems: inaccuracy in coding data, computational errors, errors in reading computer output, typographical errors by secretaries or by printers.
    Note: These errors can be very large in magnitude.

11. Variance due to extraneous factors that affect the relationship:
    Study validity will be systematically lower than true validity if incumbents differ in experience at the time they are assessed for performance.
Meta-Analysis of Correlations Corrected Individually for Artifacts

then the estimated mean and standard deviation of true effect size correlations are not corrected for the effect of that artifact. The estimates will be inaccurate to the extent that uncorrected artifacts have a substantial impact in that research domain.

Although there are 10 potentially correctable artifacts, they will not all be discussed at the same level of detail. First, sampling error is both nonsystematic and of devastating effect in narrative reviews of the literature. Thus, sampling error will be discussed first and in considerable detail. The systematic artifacts will then be considered one by one. Here, too, there will be differences in the length of the presentation. This does not mean that some are less important than others. Rather, it is a matter of mathematical redundancy. Most of the artifacts have the effect of attenuating the true correlation by a multiplicative fraction, and hence, these artifacts all have a very similar mathematical structure. Once we look at error of measurement and range variation in detail, the others are mathematically similar and, hence, can be treated more briefly. However, it is important to remember that an artifact that may have little or no effect in one research domain may have a large effect in another. For example, there are research domains where the dependent variable has never been dichotomized; hence, there need be no correction for that artifact at all. In research on employee turnover, however, almost every study dichotomizes the dependent variable, and the dichotomization is based on administrative conventions that often lead to very extreme splits. Thus, none of the artifacts in Table 3.1 can be routinely ignored, no matter how short our treatment of that artifact may be and regardless of whether it can be corrected or not.

Consider the effect of sampling error on the study correlation. At the level of the single study, sampling error is a random event. If the observed correlation is .30, then the unknown population correlation could be higher than .30 or lower than .30, and there is no way that we can know the sampling error or correct for it. However, at the level of meta-analysis, sampling error can be estimated and corrected for. Consider first the operation of averaging correlations across studies. When we average correlations, we also average the sampling errors. Thus, the sampling error in the average correlation is the average of the sampling errors in the individual correlations. For example, if we average across 30 studies with a total sample size of 2,000, then sampling error in the average correlation is about the same as if we had computed a correlation on a sample of 2,000. That is, if the total sample size is large, then there is very little sampling error in the average correlation. The variance of correlations across studies is another story. The variance of correlations is the average squared deviation of the study correlation from its mean. Squaring the deviation eliminates the sign of the sampling error and, hence, eliminates the tendency for errors to cancel themselves out in summation. Instead, sampling error causes the variance across studies to be systematically larger than the variance of population correlations that we would like to know. However, the effect of sampling error on the variance is to add a known constant to the variance, which is the sampling error variance. This constant can be subtracted from the observed variance. The difference is then an estimate of the desired variance of population correlations.

To eliminate the effect of sampling error from a meta-analysis, we must derive the distribution of population correlations from the distribution of observed correlations. That is, we would like to replace the mean and standard deviation of the
observed sample correlations by the mean and standard deviation of the population correlations. Because sampling error cancels out in the average correlation across studies, our best estimate of the mean population correlation is simply the mean of the sample correlations. However, sampling error adds to the variance of correlations across studies. Thus, we must correct the observed variance by subtracting the sampling error variance. The difference is then an estimate of the variance of population correlations across studies.

Once we have corrected the variance across studies for the effect of sampling error, it is possible to see if there is any real variance in results across studies. If there is a large amount of variance across studies, then it is possible to look for moderator variables to explain this variance. To test our hypothesized moderator variable, we break the set of studies into subsets using the moderator variable. For example, we might split the studies into those done on large corporations and those done on small businesses. We then do separate meta-analyses within each subset of studies. If we find large differences between subsets, then the hypothesized variable is indeed a moderator variable. The meta-analysis within subsets also tells us how much of the residual variance within subsets is due to sampling error and how much is real. That is, the meta-analysis tells us whether or not we need look for a second moderator variable.

Although it is pedagogically useful for us to present the search for moderator variables immediately after presenting the method for eliminating the effects of sampling error, that search is actually premature. Sampling error is only one source of artifactual variation across studies. We should eliminate other sources of variance before we look for moderator variables. Another important source of correctable variation across studies in most domains is variation in error of measurement across studies. That is, a variable such as job satisfaction can be measured in many ways. Thus, different studies will often use different measures of the independent variable or different measures of the dependent variable. Alternate measures will differ in the extent to which they are affected by error of measurement. Differences in amount of measurement error produce differences in the size of the correlations. Differences in correlations across studies due to differences in error of measurement often look like differences due to a moderator variable. Thus, we obtain a true picture of the stability of results across studies only if we eliminate the effects of measurement error. The same is true of other study design artifacts. However, measurement error is always present in every study, while other artifacts, such as dichotomization or range restriction, are sometimes present and sometimes not present.

Correctable artifacts other than sampling error are systematic rather than unsystematic in their impact on study correlations. Let us discuss error of measurement as one example of a correctable systematic artifact. At the level of the individual person, error of measurement is a random event. If Bill’s observed score is 75, then his true score could be either greater than 75 or less than 75, and there is no way of knowing which. However, when we correlate scores across persons, the random effects of error of measurement produce a systematic effect on the correlation coefficient. Error of measurement in either variable causes the correlation to be lower than it would have been with perfect measurement. We will present a formula for “attenuation” that expresses the exact extent to which the correlation
is lowered by any given amount of error of measurement. This same formula can be algebraically reversed to provide a formula for “correction for attenuation.” That is, if we know the amount of error of measurement in each variable, then we can correct the observed correlation to provide an estimate of what the correlation would have been had the variables been perfectly measured.

The amount of error of measurement in a variable is measured by a number called the reliability of the variable. The reliability is a number between 0 and 1 that measures the percentage of the observed variance that is true score variance. That is, if the reliability of the independent variable is .80, then 80% of the variance in the scores is due to the true score variation, and, by subtraction, 20% of the variance is due to variation in errors of measurement. To correct for the effect of error of measurement on the correlation, we need to know the amount of error of measurement in both variables. That is, to correct the correlation for attenuation, we need to know the reliability of both variables.

Error of measurement can be eliminated from a meta-analysis in either of two ways: at the level of single studies or at the level of averages across studies. If the reliability of each variable is known in each study, then the correlation for each study can be separately corrected for attenuation. We can then do a meta-analysis on the corrected correlations. This type of meta-analysis is the subject of this chapter. However, many studies do not report the reliability of their instruments. Thus, reliability information is often only sporadically available. Under such conditions, we can still estimate the distribution of the reliability of both the independent and the dependent variables. Given the distribution of observed correlations, the distribution of the reliability of the independent variables, and the distribution of the reliability of the dependent variable, it is possible to use special formulas to correct the meta-analysis to eliminate the effects of error of measurement. Meta-analysis based on such artifact distributions is the subject of the next chapter.

If each individual correlation is corrected for attenuation, then the meta-analysis formulas will differ slightly from the formulas for meta-analysis on uncorrected correlations. The average corrected correlation estimates the average population correlation between true scores. The observed variance of the corrected correlations can be corrected for sampling error simply by subtracting a constant—the sampling error variance. However, the sampling error in a corrected correlation is larger than the sampling error in an uncorrected correlation. Therefore, a different formula must be used to compute the sampling error variance for corrected correlations.

The other correctable artifact most studied in psychometric theory (and in personnel selection research) is range restriction (though our formulas also handle the case of range enhancement or range variation). In many contexts, the standard deviation of the independent variable is approximately the same across studies (i.e., is the same to within sampling error—which will produce some variation in standard deviations). In such cases, the meta-analysis need not correct for range variation. However, if the standard deviation of the independent variable differs radically from study to study, then there will be corresponding differences in the correlation from study to study. These differences across studies will look like differences produced by a moderator variable. Thus, if there are big differences
in the standard deviation of the independent variable across studies, then a true picture of the stability of results will appear only if the effects of range variation are eliminated. To do this, we compute the value that the correlation would have had had the study been done on a population with some reference level of variance on the independent variable.

Range deviation can be corrected at the level of the single study. If we know the standard deviation of the independent variable in the study, and if we know the standard deviation in the reference population, then there are range correction formulas that will produce an estimate of what the correlation would have been had the standard deviation of the study population been equal to the standard deviation of the reference population. As noted in the previous chapter, these correction procedures are different for direct and indirect range variation. If we correct a correlation for range departure, then the corrected correlation will have a different amount of sampling error than an uncorrected correlation. Therefore, a meta-analysis on corrected correlations must use a different formula for the sampling error variance.

In an ideal research review, we would have complete information about artifacts on each study. For each study, we would know the extent of range departure and the reliabilities of both variables. We could then correct each correlation for both range departure and error of measurement. We would then do a meta-analysis of the fully corrected correlations.

Discussion of other correctable artifacts will be taken up when the needed mathematical tools have been presented. Skipping these artifacts at this time is by no means intended to imply that they are less important. For example, in turnover studies, dichotomization has an even larger attenuating effect on study correlations than does error of measurement.

The remainder of this chapter is presented in four main sections. First, we give a complete treatment of meta-analysis with correction for sampling error only. Second, we present a detailed treatment of error of measurement and range departure, both in terms of the corrections for single studies and in terms of the effect of the correction on sampling error. Third, we describe a more abbreviated treatment of each of the other correctable artifacts. Fourth, we present meta-analysis for the case of individually corrected correlations, that is, meta-analysis as it is done when full information is available on the artifact values in each study. Examples of published studies of this sort include Carlson, Scullen, Schmidt, Rothstein, and Erwin (1999), Judge, Thorensen, Bono, and Patton (2001), and Rothstein, Schmidt, Erwin, Owens, and Sparks (1990).

At present, we know of no research domain where the information has been available to correct for every one of the correctable artifacts listed in Table 3.1. Some meta-analysis methods (e.g., Glass, McGaw, & Smith, 1981) do not even correct for sampling error; they take each study outcome at face value (see Chapter 11). In personnel selection research, correction is typically made only for sampling error, error of measurement, and range restriction. Most current meta-analyses have made no correction for dichotomization or imperfect construct validity. Furthermore, there will probably be more correctable artifacts defined and quantified over the coming years. And even if all correctable artifacts were corrected, there is still reporting error and other data errors.
It is important to keep in mind that even a fully corrected meta-analysis will not correct for all artifacts. Even after correction, the remaining variation across studies should be viewed with skepticism (see Chapters 2 and 5). Small residual variance is probably due to uncorrected artifacts rather than to a real moderator variable.

**Bare-Bones Meta-Analysis: Correcting for Sampling Error Only**

We will now present a detailed discussion of sampling error. To keep the presentation simple, we will ignore other artifacts. The resulting presentation is thus written as if the study population correlations were free of other artifacts. In later sections, we will discuss the relationship between sampling error and other artifacts. This section also presents the mathematics of a meta-analysis in which sampling error is the only artifact corrected. Alas, there are those who do not believe that other artifacts exist. There are also those who believe that if there are artifacts in the original studies, then those same artifacts should be reflected in the meta-analysis. They believe that the purpose of meta-analysis is only to describe observed results and that meta-analysis should not correct for known problems in research design (see Chapter 11). However, most scientists believe that the goal of cumulative research is to produce better answers than can be obtained in isolated studies (Rubin, 1990). From that point of view, the purpose of meta-analysis is to estimate the relationships that would have been observed if studies had been conducted perfectly (Rubin, 1990), that is, to estimate construct-level relationships (see Chapters 1, 11, and 14). Given this purpose for meta-analysis, a meta-analysis that does not correct for as many artifacts as possible is an unfinished meta-analysis. We hold that view. If a meta-analysis corrects only for sampling error, then it is the mathematical equivalent of the ostrich with its head in the sand: It is a pretense that if we ignore other artifacts then their effects on study outcomes will go away.

**Estimation of Sampling Error**

If the population correlation is assumed to be constant over studies, then the best estimate of that correlation is not the simple mean across studies but a weighted average in which each correlation is weighted by the number of persons in that study. Thus, the best estimate of the population correlation is

$$
\bar{r} = \frac{\sum N_i r_i}{\sum N_i}
$$

where $r_i$ is the correlation in study $i$ and $N_i$ is the number of persons in study $i$. The corresponding variance across studies is not the usual sample variance, but the frequency-weighted average squared error

$$
s^2_r = \frac{\sum [N_i (r_i - \bar{r})^2]}{\sum N}
$$
Two questions are often asked about this procedure. First, is the weighted average always better than the simple average? Hunter and Schmidt (1987a) presented a detailed discussion of this. Their analysis showed that it is a very rare case in which an unweighted analysis would be better. Second, why do we not transform the correlations to Fisher $z$ form for the cumulative analysis? The answer is that the Fisher $z$ transformation produces an estimate of the mean correlation that is upwardly biased and less accurate than an analysis using untransformed correlations (see Chapter 5; also, Hall & Brannick, 2002; Hunter, Schmidt, & Coggin, 1996).

The frequency-weighted average gives greater weight to large studies than to small studies. If there is no variance in population correlations across studies, then the weighting always improves accuracy. If the variance of population correlations is small, then the weighted average is also always better. If the variance of population correlations across studies is large, then as long as sample size is not correlated with the size of the population correlation, the weighted average will again be superior. That leaves one case in which the weighted average could prove troublesome. For example, in one meta-analysis, we found 13 studies on the validity of bio-data in predicting job success. One of these studies was done by an insurance consortium with a sample size of 15,000. The other 12 studies were done with sample sizes of 500 or less. The weighted average will give the single insurance study over 30 times the weight given to any other study. Suppose that the insurance study were deviant in some way. The meta-analysis might then be almost entirely defined by that deviant study. In a situation such as this, we recommend two analyses: a first analysis with the large-sample study included and a second analysis with the large-sample study left out. We have not yet had to figure out what to do should the two analyses show a major discrepancy. (In our case, they did not.) (See the related discussion of fixed- vs. random-effects meta-analysis models in Chapters 5 and 9. In particular, see the discussion of methods of weighting studies in random vs. fixed models of meta-analysis.)

What about the Fisher $z$ transformation? In our original work, we carried out calculations with and without use of Fisher’s $z$. For preliminary calculations done by hand, we averaged the correlations themselves, but on the computer we used what we thought to be the superior Fisher $z$ transformation. For several years, we noticed no difference in the results for these two analyses, but in our validity generalization study for computer programmers (Schmidt, Gast-Rosenberg, & Hunter, 1980), the difference was notable. The average validity using the Fisher $z$ transformation was larger (by about .03) than the average validity when correlations were averaged without this transformation. Careful checking of the mathematics then showed that it is the Fisher transformation that is biased. The Fisher $z$ transformation gives larger weights to large correlations than to small ones, hence the positive bias. This problem is discussed in more detail in Chapter 5. (See “Accuracy of Different Random-Effects Models.”)

Although the Fisher $z$ transformation produces an upward bias when it is used in averaging correlations, the transformation does serve its original purpose quite well. The original purpose was not to create a method for averaging correlations. Fisher’s purpose was to create a transformation of the correlation for which the standard error (and, therefore, confidence intervals) would depend solely on the
sample size and not on the size of the statistic. The standard error of the Fisher $z$ statistic is $1/(N - 3)^{1/2}$, and so this goal was achieved. This means that, unlike the case for the correlation, it is unnecessary to have an estimate of the population value to compute the standard error and confidence intervals.

There has been considerable confusion in the literature produced by the fact that there is a slight bias in the correlation coefficient (a bias that can be corrected, as noted in Chapter 2 and below). There is a widespread false belief that the Fisher $z$ eliminates that bias. The fact is that the Fisher $z$ replaces a small underestimation, or negative bias, by a typically small overestimation, or positive bias, a bias that is always greater in absolute value than the bias in the untransformed correlation. This bias is especially large if there is variation in the population correlations across studies (Hunter et al., 1996; Schmidt & Hunter, 2003). In this case, the bias in Fisher’s $z$ can cause estimates of the mean correlation to be biased upward by substantial amounts (Field, 2001; Hall & Brannick, 2002; Schmidt & Hunter, 2003). This appears to be the reason that the random-effects meta-analysis methods of Hedges and Olkin (1985) overestimate the mean correlation (Field, 2001; Hall & Brannick, 2002). (See Chapters 5 and 9 for more detail on this issue.) It appears that meta-analysis is never made more accurate by using the Fisher $z$ transformation and can be made substantially less accurate under certain conditions.

It is well-known that the variance of any variable can be computed using either $N$ or $N - 1$ in the denominator. The formulation in Equation (3.2) corresponds to use of $N$ in the denominator. The advantage of using $N$ in the denominator is that it leads to a more accurate estimate of the variance in the sense of statistical efficiency. That is, root mean square error is lower, an important consideration in meta-analysis. Use of $N - 1$ instead of $N$ leads to a slight reduction in bias but at the expense of a reduction in overall accuracy.

**Correcting the Variance for Sampling Error and a Worked Example**

Consider the variation in correlations across similar studies on a research question. The observed variance $s^2_r$ is a confounding of two things: variation in population correlations (if there is any) and variation in sample correlations produced by sampling error. Thus, an estimate of the variance in population correlations can be obtained only by correcting the observed variance $s^2_r$ for sampling error. The following mathematics shows that sampling error across studies behaves like error of measurement across persons and that the resulting formulas are comparable to standard formulas in classical measurement theory (reliability theory).

We begin with a treatment of sampling error in an isolated study. In an isolated study, the correlation is based on a sample: a specific sample of the population of people who might have been in that place at that time, a sample of the random processes in each person’s head that generate error of measurement in test responses or supervisor ratings and so on, and a sample of time variation in person and situational parameters. This is represented in statistics by noting that the observed correlation is a sample from a population distribution of correlation values that
might have been observed if the study were replicated except for the random factors. These replications are hypothetical in the sense that a real study usually contains only one such sample. However, the replications are not hypothetical in that they do represent real variation. There have been thousands of actual experiments testing the theory of statistical sampling error, and all have verified that theory. Sampling error in an isolated study is unobservable but present nonetheless.

For any study, then, there is a real population correlation $\rho$ (which is usually unknown) that can be compared to the study correlation $r$. The difference between them is the sampling error, which we will denote by $e$. That is, we define sampling error $e$ by the formula

$$ e = r - \rho $$

or

$$ r = \rho + e. $$

The distribution of observed correlations for the (usually hypothetical) replications of the study is always centered about the population correlation $\rho$, although the sampling error varies randomly. If we ignore the small bias in the correlation coefficient (or if we correct for it as discussed later), then the average sampling error will be 0 and the standard deviation of the sampling error will depend on the sample size. The sampling error in a particular correlation can never be changed (and, in particular, is not changed when a statistical significance test is run). On the other hand, if replications of the study could be done, then sampling error could thus be reduced. Because the average error is 0, the replicated correlations can be averaged, and the average correlation is closer to the population correlation than the individual correlations. The sampling error in the average correlation is the average of the individual sampling errors and is thus much closer to 0 than the typical single sampling error. The average correlation has smaller sampling error in much the same way as a correlation based on a larger sample size. Thus, replicating studies can potentially solve the problem of sampling error.

Whereas replication is not possible in most individual studies, replication does take place across different studies. Consider the ideal special case: a meta-analysis for a research domain in which there is no variation in the population correlations across studies and in which all studies are done with the same sample size. This case is mathematically identical to the hypothetical replications that form the basis of the statistics of isolated correlations. In particular, the average correlation across studies would have greatly reduced sampling error. In the case in which the population correlation varies from one study to another, the replication is more complicated, but the principle is the same. Replication of sampling error across studies enables us to use averaging to reduce the impact of sampling error. If the number of studies is large, the impact of sampling error can be virtually eliminated.

How big is the typical sampling error? Because sampling error has a mean of 0, the mean sampling error does not measure the size of the sampling error. That is, because a negative error of $-.10$ is just as bad as a positive error of $+.10$, it is the absolute value of the error that counts. To assess the size of errors without the algebraic sign, the common statistical practice is to square the errors. The
average squared error is the variance, and the square root of that is the standard deviation of the errors. It is the standard deviation of sampling error that is the best representation of the size of errors. Consider, then, the isolated study. In the common case in which the underlying distribution is the bivariate normal distribution, the standard deviation of the sampling error is given by

$$\sigma_e = \frac{(1 - \rho^2)}{\sqrt{(N - 1)}}$$

where \(N\) is the sample size. Technically, our use of this formula throughout this book is tantamount to the assumption that all studies are done in contexts where the independent and dependent variables have a bivariate normal distribution, but the statistics literature has found this formula to be fairly robust in the face of departures from normality. However, under conditions of range restriction, this formula underestimates sampling error variance to some extent. That is, actual sampling variance is greater than the value predicted by the formula, leading to undercorrections for sampling error variance. Using computer simulation, Millsap (1989) showed this for direct range restriction, and Aguinis and Whitehead (1997) showed it for indirect range restriction. This means that corrections for sampling error variance are undercorrections, leading to overestimates of SD\(_e\), particularly in validity generalization studies. This could create the appearance of a moderator where none exists. However, in research areas where there is no range restriction, this problem does not occur.

The effect of averaging across replications is dramatic in terms of sampling error variance. If the sample size in each replication is \(N\) and the number of studies is \(K\), then the sampling error variance in the average of \(K\) correlations is the variance of the average error \(e\), that is,

$$\text{Var}(\bar{e}) = \frac{\text{Var}(e)}{K} \quad (3.3)$$

In other words, the effect of averaging across \(K\) studies is to divide the sampling error variance by \(K\). Because the total sample size in \(K\) studies is \(K\) times the sample size in a single study, this means that to increase the sample size by a factor of \(K\) is to reduce the sampling error variance by a factor of \(K\). This is exactly the same rule as that for increasing the sample size of a single study. Thus, replication can reduce sampling error in the same way as using larger samples.

In practice, the effect of increasing sample size is not quite as impressive as the previous formula would suggest. Unfortunately, it is not the variance but the standard deviation (standard error) that counts. The standard deviation of the average error is divided only by the square root of the number of studies. Thus, to cut the standard error in half, we must average four studies rather than two studies. This is important in judging the number of missing studies in a meta-analysis when that number is small. For example, if an investigator randomly misses 10 out of 100 potential studies, the sampling error variance is increased by 10%, but the sampling error standard error is increased by only 5%. Thus, missing a few studies randomly usually does not reduce the accuracy of a meta-analysis by nearly as much as might be supposed.

We come now to the meta-analysis of correlations taken from different studies. The power of meta-analysis to reduce the problem of sampling error lies in the
fact that sampling errors are replicated across studies. The ultimate statistical error in meta-analysis will depend on two factors: the size of the average sample size for the individual studies and the number of studies in the meta-analysis. For the mean correlation, it is the total sample size that determines the error in the meta-analysis. For the estimate of the variance of correlations, the computations are more complicated but the principle is similar.

Let the subscript \( i \) denote the study number. Then the error variable \( e_i \) represents the sampling error in the sample correlation in study \( i \); that is, we define \( e_i \) by

\[
   r_i = \rho_i + e_i
\]

Then the mean of the error within hypothetical replications is 0. The average error across studies is

\[
   E(e_i) = 0
\]

The variance across hypothetical replications of any one study is denoted by

\[
   \sigma^2_{e_i} = \frac{(1 - \rho^2_i)^2}{N_i - 1}
\]

(3.4)

In going across studies, this hypothetical and unobserved variation becomes real and potentially observable variation. This is similar to the case of observing a sample of people or scores from a distribution of people or scores. It differs from the usual case in statistics because the sampling error variance differs from one study to the next. Nonetheless, the critical fact is that the variance of sampling error becomes visible across studies. The formula

\[
   r_i = \rho_i + e_i
\]

is analogous to the true score, error score formula from measurement error theory:

\[
   X_p = T_p + e_p
\]

where \( X_p \) and \( T_p \) are the observed and true scores for person \( p \). In particular, sampling error (signed sampling error, not sampling error variance) is unrelated to population values across studies (cf. Schmidt, Hunter, & Raju, 1988). Thus, if we calculate a variance across studies, then the variance of sample correlations is the sum of the variance in population correlations and the variance due to sampling error, that is,

\[
   \sigma^2_r = \sigma^2_p + \sigma^2_e
\]

The implication of this formula is that the variance of observed correlations is larger than the variance of population correlations, often much larger. The reason it is larger is because squared sampling errors are always positive and do not cancel out when averaged. Thus, the average squared deviation of observed correlations is systematically larger than the average squared deviation of population correlations because sampling error makes a systematically positive contribution to the squared deviation.
The formula $\sigma^2 = \sigma^2_\rho + \sigma^2_e$ has three variances. If any two of the variances are known, the third can be computed. In particular, if the sampling error variance $\sigma^2_e$ were known, then the desired variance of population correlations would be

$$\sigma^2_\rho = \sigma^2_r - \sigma^2_e$$

Of these three variances, only the variance of observed correlations is estimated using a conventional variance, that is, average squared deviation of given numbers. If we knew the value of the sampling error variance, then it would not matter that we could not compute it as a conventional variance. The fact is that the sampling error variance need not be given as an empirical number; it is given by statistical formula. The sampling error variance across studies is just the average of the sampling error variances within studies. If study correlations are weighted by sample size $N_i$, then

$$\sigma^2_e = \text{Ave} \sigma^2_{ei} = \frac{\sum [N_i \sigma^2_{ei}]}{\sum N_i} = \frac{\sum [N_i [(1 - \rho^2_i)^2]/[N_i - 1]]}{\sum N_i}$$

(3.5)

This formula is used in our Windows-based meta-analysis computer program package. (See the Appendix for a description and availability.) However, Law, Schmidt, and Hunter (1994b) and Hunter and Schmidt (1994b) showed via computer simulation that this formula is more accurate if $\rho_i$ is estimated by $\bar{r}$. That is, in computing the sampling error variance for each individual study, $\bar{r}$ is used in place of $r_i$. This modification is used in our computer programs. (See Chapter 5 for further discussion of this.)

Approximation formulas are available for use in hand calculations of meta-analysis. In Equation (3.5), the fraction $N_i/(N_i - 1)$ is close to unity. If we take this fraction as unity, and we use the approximation that average $(\rho^2) \cong (\text{average} \, \rho)^2$, then we have the almost perfect approximation

$$\sigma^2_e = \frac{(1 - \bar{r}^2)^2 K}{T}$$

(3.6)

where $K$ is the number of studies and $T = \sum N_i$ is the total sample size. The corresponding estimate of the variance of population correlations is thus

$$\text{est} \, \sigma^2_\rho = \sigma^2_r - \sigma^2_e = \sigma^2_r - \frac{(1 - \bar{r}^2)^2 K}{T}$$

There is an even better estimate of the sampling error variance. Consider the special case in which all studies have the same sample size $N$. The ratio $N_i/(N_i - 1)$ is then simply the constant $N/(N - 1)$, which factors out of the summation. We then have

$$\sigma^2_e = \text{Ave} \frac{(1 - \rho_i^2)^2}{N - 1}$$

If we estimate $\rho_i$ by the average correlation across studies (Hunter & Schmidt, 1994b; Law et al., 1994b), we have the approximation

$$\sigma^2_e = \frac{(1 - \bar{r}^2)^2}{(N - 1)}$$

(3.7)
This formula is exactly analogous to the formula for sampling error in the single study

\[ \sigma_e^2 = \frac{(1 - \rho^2)^2}{(N - 1)} \]

The relationship between the first approximation and the second approximation stems from the fact that

\[ K/T = 11/(T/K)(T/K) = 1/N \]

That is, the previous approximation with \( K \) in the numerator and \( T \) in the denominator is equivalent to having average sample size in the denominator. Thus, the improvement is to use \( N - 1 \) instead of \( N \), a small change for typical sample sizes of 100, but a noticeable improvement in those unusual meta-analyses conducted in research areas with very small sample sizes. For example, psychotherapy studies have a typical sample size of about 20.

Consider again the typical case where sample size varies from one study to the next. Let the average sample size be denoted by \( \bar{N} \), that is,

\[ \bar{N} = T/K \]

Our first approximation could be written as

\[ \sigma_e^2 = \frac{(1 - \bar{r}^2)^2}{\bar{N}} \]

while the second improved approximation is

\[ \sigma_e^2 = \frac{(1 - \bar{r}^2)^2}{(\bar{N} - 1)} \]

Mathematical work not presented here (Hunter & Schmidt, 1987a) shows that the second approximation is accurate both when the population correlations are all the same (Hunter & Schmidt, 1994b)—when it is optimal—and also when there is variation in population correlations (when complicated weighting schemes based on knowledge of the distribution of sample size—not usually known in a meta-analysis—would improve the estimate slightly; Law et al., 1994b). The corresponding estimate of the variance of population correlations is

\[ \sigma_\rho^2 = \sigma_r^2 - \sigma_e^2 = \sigma_r^2 - (1 - \bar{r}^2)^2/(\bar{N} - 1) \]

This equation allows an empirical test of the hypothesis \( \sigma_\rho^2 = 0 \).

**An Example: Socioeconomic Status and Police Performance**

Bouchard (1776, 1860, 1914, 1941) postulated that differences in upbringing would produce differences in response to power over other people. His theory was that because lower-class parents obtain obedience by beating their children to a pulp while middle-class parents threaten them with loss of love, lower-class children would grow into adults who are more likely themselves to use physical force to gain compliance. He tested his theory by looking at the relationship between socioeconomic status of origin of police officers and brutality in police
departments. His independent measure was socioeconomic status measured in terms of six classes, ranging from 1 = upper, upper class to 6 = lower, lower class. His brutality measure was the number of complaints divided by the number of years employed. Only patrol officers were considered in the correlations, which are shown in Table 3.2. The meta-analysis of these data is as follows:

\[ \bar{r} = \frac{100(.34) + 100(.16) + 50(.12) + 50(.38)}{100 + 100 + 50 + 50} = \frac{75.00}{300} = .25 \]

\[ \sigma^2_r = \frac{100(.34 - .25)^2 + 100(.16 - .25)^2 + 50(.12 - .25)^2 + 50(.38 - .25)^2}{100 + 100 + 50 + 50} = \frac{3.31}{300} = .011033 \]

The average sample size is

\[ \bar{N} = \frac{T}{K} = \frac{300}{4} = 75 \]

Thus, the sampling error variance is estimated to be

\[ \sigma^2_e = \frac{(1 - \bar{r}^2)^2}{(\bar{N} - 1)} = \frac{(1 - .25^2)^2}{74} = .011877 \]

The estimate of the variance of population correlations is thus

\[ \sigma^2_\rho = \sigma^2_r - \sigma^2_e = \sigma^2_r - \frac{(1 - \bar{r}^2)^2}{(\bar{N} - 1)} = .011033 - .011877 = -.000844. \]

Because the estimated variance is negative, the estimated standard deviation is 0, that is,

\[ \sigma_\rho = 0 \]

Some readers have been bothered by this example. They ask, “How can a variance be negative, even if only −.0008?” The answer is that the estimated variance of population correlations is not computed as a conventional variance, that is, the average squared deviation of given numbers. Rather, it is computed as the difference between the given variance of observed correlations and the statistically given sampling error variance. Although there is little error in the statistically given sampling error variance, the variance of observed correlations is a sample estimate. Unless the number of studies is infinite, there will be some error in that empirical estimate. If the population difference is 0, then error will cause the estimated difference to be positive or negative with probability one-half. Thus, in our case, sampling error caused the variance of observed correlations to differ slightly from the expected value, and that error caused the estimating difference to be negative. There is no logical contradiction here. In analysis of variance and in Cronbach’s generalizability theory, estimation of components of variance using expected mean square formulas also produces negative observed estimates for similar reasons. Such estimates are always taken as 0. This question is discussed further in Chapter 9. (See discussion of second-order sampling error.)
Consider the empirical meaning of our results. Bouchard claimed that his results varied dramatically from city to city. His explanation was that Washington, D.C., and Richmond are southern cities, and southern hospitality is so strong that it reduces the incidence of brutality in the lower classes and, hence, reduces the correlation in those cities. However, our analysis shows that all the variation in his results is due to sampling error and that the correlation is always .25.

Moderator Variables Analyzed by Grouping the Data and a Worked Example

A moderator variable is a variable that causes differences in the correlation between two other variables. For example, in the police brutality study discussed previously, Bouchard postulated that geographic region (North vs. South) would be a moderator variable for the relationship between socioeconomic status and brutality. If there is true variation in results across studies, then there must be such a moderator variable (or possibly more than one) to account for such variance. On the other hand, if the analysis shows that the variation in results is due to sampling error, then any apparent moderating effect is due to capitalization on sampling error. This was the case in Bouchard’s work.

If the corrected standard deviation suggests substantial variation in population correlations across studies, then a moderator variable derived from a theory or hypothesis can be used to group the observed correlations into subsets. Within each subset, we can calculate a mean, a variance, and a variance corrected for sampling error. A moderator variable will reveal itself in two ways: (1) the average correlation will vary from subset to subset, and (2) the corrected variance will average lower in the subsets than for the data as a whole. These two facts are mathematically dependent. By a theorem in analysis of variance, we know that the total variance is the mean of the subset variances plus the variance of the subset means. Thus, the mean uncorrected within-subset variance must decrease to exactly the extent that the subset means differ from one another. This means that if the average correlation varies across subsets, then the average standard deviation of the subsets must be less than the standard deviation in the combined data set.

An Example: Police Brutality in Transylvania

In order to justify a European sabbatical, Hackman (1978) argued that Bouchard’s work on police brutality needed a cross-cultural replication. So he
gathered data in four cities in Transylvania, carefully replicating Bouchard’s measurement on socioeconomic status and brutality. His data are given along with Bouchard’s in Table 3.3.

**Analysis of the Whole Set**

\[
\bar{r} = \frac{100(0.34) + \cdots + 100(0.19) + \cdots + 50(0.23)}{100 + \cdots + 100 + \cdots + 50} = \frac{105.00}{600} = 0.175
\]

\[
\sigma_r^2 = \frac{100(0.34 - 0.175)^2 + \cdots + 50(0.23 - 0.175)^2}{100 + \cdots + 50} = \frac{9.995}{600} = 0.016658
\]

\[N = T/K = 600/8 = 75\]

\[
\sigma_e^2 = \frac{(1 - 0.175^2)^2}{74} = 0.12698
\]

\[
\sigma_p^2 = 0.016658 - 0.012698 = 0.00396
\]

\[
\sigma = 0.063
\]

The corrected standard deviation of 0.063 can be compared with the mean of 0.175 : 0.175/0.063 = 2.78. That is, the mean correlation is nearly 2.8 standard deviations above 0. Thus, if the study population correlations are normally distributed, the probability of a zero or below-zero correlation is virtually nil. So the qualitative nature of the relationship is clear: The population correlation is positive in all studies.

However, the variation is not trivial in amount relative to the mean. This suggests a search for moderator variables. The moderator analysis is as follows:

<table>
<thead>
<tr>
<th>United States</th>
<th>Transylvania</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\bar{r} = .25)</td>
<td>(\bar{r} = .10)</td>
</tr>
<tr>
<td>(\sigma_r^2 = .011033)</td>
<td>(\sigma_r^2 = .011033)</td>
</tr>
<tr>
<td>(\sigma_e^2 = .011877)</td>
<td>(\sigma_e^2 = .013245)</td>
</tr>
<tr>
<td>(\sigma_p^2 = -.000844)</td>
<td>(\sigma_p^2 = -.002212)</td>
</tr>
<tr>
<td>(\sigma = 0)</td>
<td>(\sigma = 0)</td>
</tr>
</tbody>
</table>

Analysis of the subsets shows a substantial difference in mean correlations, \(\bar{r} = .25\) in the United States and \(\bar{r} = .10\) in Transylvania. The corrected standard deviations reveal that there is no variation in results within the two countries.

In this case, there was only one moderator. When there are multiple moderators, the moderators may be correlated, and hence, they will be confounded if examined sequentially one at a time. In these cases, it is important to conduct moderator analyses hierarchically to avoid such confounding. (See Chapter 9.)

Hackman explained the difference between the two countries by noting that vampires in the United States live quiet, contented lives working for the Red Cross, while vampires in Transylvania must still get their blood by tracking down and killing live victims. Vampires in Transylvania resent their low station in life and focus their efforts on people of high status, whom they envy. Middle-class policemen who work at night are particularly vulnerable. Thus, there is less variance in social class among the policemen in Transylvania, and this restriction in
Table 3.3 Correlations between socioeconomic status and police brutality (U.S. and Transylvania)

<table>
<thead>
<tr>
<th>Investigator</th>
<th>Location</th>
<th>Sample Size</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bouchard</td>
<td>Philadelphia</td>
<td>100</td>
<td>.34*</td>
</tr>
<tr>
<td>Bouchard</td>
<td>Richmond, VA</td>
<td>100</td>
<td>.16</td>
</tr>
<tr>
<td>Bouchard</td>
<td>Washington, DC</td>
<td>50</td>
<td>.12</td>
</tr>
<tr>
<td>Bouchard</td>
<td>Pearl Harbor</td>
<td>50</td>
<td>.38*</td>
</tr>
<tr>
<td>Hackman</td>
<td>Brasov</td>
<td>100</td>
<td>.19</td>
</tr>
<tr>
<td>Hackman</td>
<td>Targul-Ocna</td>
<td>100</td>
<td>.01</td>
</tr>
<tr>
<td>Hackman</td>
<td>Hunedoara</td>
<td>50</td>
<td>-.03</td>
</tr>
<tr>
<td>Hackman</td>
<td>Lupeni</td>
<td>50</td>
<td>.23</td>
</tr>
</tbody>
</table>

*Significant at the .05 level.

range reduces the correlation. According to this hypothesis, correcting for range restriction would make the two mean correlations equal (at .25). Later in this chapter, we will examine range corrections that can be used to test this hypothesis.

After a heated exchange at the Academy of Management convention, Bouchard bared his fangs and showed Hackman that American vampires can still be a pain in the neck. Bouchard then noted that the difference in the results reflected the fact that his studies were done at times when the country was going to war. This increase in aggressive excitement increased the general level and the variance of brutality and, thus, increased its reliability of measurement and, hence, the level of correlation. According to this hypothesis, correcting for measurement error in the brutality measure would reveal that the two mean correlations are really equal at the level of true scores. This hypothesis can be tested using the corrections for measurement error discussed later in this chapter.

Correcting Feature Correlations for Sampling Error and a Worked Example

Suppose some study feature is coded as a quantitative variable $y$. Then that feature can be correlated with the outcome statistic across studies. For example, if correlations between dependency and school achievement varied as a function of the age of the child, then we might code average age in study $i$ as $y_i$. We could then correlate age of children with size of correlation across studies. An example of this method is given by Schwab, Olian-Gottlieb, and Heneman (1979). However, such a correlation across studies is a confounding of the correlation for population values with $y$ and the noncorrelation of the sampling error with $y$. This is directly analogous to the role of error of measurement in attenuating correlations based on imperfectly measured variables. Thus, the observed correlation across studies will be smaller than would be the case had there been no sampling error in the correlations.

To avoid confusion between the basic statistic $r$, which is the correlation over persons within a study, and correlations between $r$ and study features over studies, the correlations over studies will be denoted by “Cor.” For example, the correlation
Meta-Analysis of Correlations Corrected Individually for Artifacts

between the correlation \( r \) and the study feature \( y \) across studies will be denoted \( \text{Cor}(r, y) \). This is the observed correlation across studies, but the desired correlation across studies is that for population correlations, \( \rho \). The desired correlation across studies is \( \text{Cor}(\rho, y) \). Starting from the formula \( r_i = \rho_i + \epsilon_i \), we calculate a covariance over studies and use the principle of additivity of covariances to produce

\[
\sigma_{ry} = \sigma_{py} + \sigma_{ey} = \sigma_{py} + 0 = \sigma_{py}
\]

If this covariance across studies is divided by standard deviations across studies, then we have

\[
\text{Cor}(r, y) = \frac{\sigma_{ry}}{\sigma_r \sigma_y} = \frac{\sigma_{py}}{\sigma_r \sigma_y} + \frac{\sigma_{ey}}{\sigma_r \sigma_y} = \frac{\sigma_{py}}{\sigma_r \sigma_y} + 0 = \frac{\sigma_{py}}{\sigma_r}
\]

However, the covariance of \( r_i \) with \( \rho_i \) is

\[
\sigma_{rp} = \sigma_{pp} + \sigma_{ep} = \sigma_{pp} + 0 = \sigma_{p}^2
\]

Hence, the correlation across studies is

\[
\text{Cor}(r, \rho) = \frac{\sigma_{rp}}{\sigma_r \sigma_\rho} = \frac{\sigma_{p}^2}{\sigma_r \sigma_\rho} = \frac{\sigma_\rho}{\sigma_r}
\]

Thus, the observed correlation across studies is the product of two other correlations, the desired correlation and reliability-like correlation:

\[
\text{Cor}(r, y) = \text{Cor}(\rho, y) \text{Cor}(r, \rho)
\]

The desired correlation is then the ratio

\[
\text{Cor}(\rho, y) = \frac{\text{Cor}(r, y)}{\text{Cor}(r, \rho)}
\]

which is precisely the formula for correction for attenuation due to error of measurement if there is measurement error in one variable only. What is the correlation between \( r \) and \( \rho \) over studies? We have the variance of \( r \) as estimated by \( \sigma_r^2 \). We need only the variance of \( \rho \), which was estimated in the previous section of this chapter. Thus, the “reliability” needed for use in the attenuation formula is given by

\[
\text{Reliability of } r = \frac{[\text{Cor}(r, \rho)]^2}{\sigma_r^2} = \frac{\sigma_\rho^2 - (1 - \bar{r}^2)}/(\bar{N} - 1)
\]

Hence,

\[
\text{Cor}(\rho, y) = \frac{\text{Cor}(r, y)}{[\sigma_r^2 - (1 - \bar{r}^2)/\bar{N} - 1]/[\sigma_r^2]^{1/2}}
\]
An Example: The Tibetan Employment Service

Officials in the Tibetan Employment Service have been using a cognitive ability test for some years to steer people into various jobs. Although they have relied on content validity for such assignments, they have also been gathering criterion-related validity data to test their content validity system. In their content validity system, test development analysts rate each occupation for the extent to which it requires high cognitive ability, with ratings from 1 = low to 3 = high. They have concurrent validity studies on six occupations chosen to stratify the full range of the content validity continuum. These data are shown in Table 3.4. The analysis is as follows:

\[
\bar{r} = .30 \\
\sigma^2_r = .048333 \\
\bar{N} = T/K = 600/6 = 100 \\
\sigma^2_N = .008365 \\
\sigma^2_\rho = .048333 - .008365 = .039968 \\
\sigma_\rho = .20 \\
\text{Rel}(r) = \frac{\sigma^2_\rho}{\sigma^2_r} = \frac{.0400}{.0483} = .83
\]

The percentage of variance due to sampling error is 17%. Let \( y_i \) be the cognitive rating of the \( i \)th occupation. Then

\[
\text{Cor}(r, y) = .72 \\
\text{Cor}(\rho, y) = \frac{.72}{\sqrt{.83}} = .79
\]

The study found very large variation in validity, even after correction for sampling error. The correlation was .72 between rating and observed correlation and rose to .79 after correction for sampling error. In this case, only 17% of the variance of the correlations was due to artifacts, so the reliability of the study correlations was .83 (i.e., \( 1 - .17 \)). Ordinarily, reliability would be much lower and, hence, the correction would be much larger. For example, if 70% of the variance were due to artifacts, then the reliability would be only \( 1 - .70 = .30 \). The correction factor would be \( 1/(.30)^{1/2} = 1.83 \).

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Cognitive Rating</th>
<th>Validity (Correlation)</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monastery Abbot</td>
<td>3</td>
<td>.45*</td>
<td>100</td>
</tr>
<tr>
<td>Magistrate</td>
<td>3</td>
<td>.55*</td>
<td>100</td>
</tr>
<tr>
<td>Holy Man</td>
<td>2</td>
<td>.05</td>
<td>100</td>
</tr>
<tr>
<td>Farmer</td>
<td>2</td>
<td>.55*</td>
<td>100</td>
</tr>
<tr>
<td>Bandit</td>
<td>1</td>
<td>.10</td>
<td>100</td>
</tr>
<tr>
<td>Yak Chip Collector</td>
<td>1</td>
<td>.10</td>
<td>100</td>
</tr>
</tbody>
</table>

*Significant at the .05 level.
In this example, we have corrected only for unreliability in the dependent variable (the validity correlations). However, as in all studies, the independent variable also contains measurement error, and so we should also correct for unreliability in the ratings of the cognitive requirements of the jobs, as pointed out by Orwin and Cordray (1985). Suppose in this particular case, the reliability of these ratings was found to be .75. Then the true score correlation between ratings and test validity would be

\[
\text{Cor}(\rho, y) = \frac{.72}{\sqrt{.83 (.75)}} = .91
\]

So it is apparent that in a real analysis of this sort, it is critical to correct for measurement error in both measures (Orwin & Cordray, 1985). Otherwise, construct-level relationships will be underestimated.

### Artifacts Other Than Sampling Error

#### Error of Measurement and Correction for Attenuation

Variables in science are never perfectly measured. Indeed, sometimes the measurement is very crude. Since the late 1890s, we have known that the error of measurement attenuates the correlation coefficient. That is, error of measurement systematically lowers the correlation between measures in comparison to the correlation between the variables themselves. This systematic error is then exaggerated by the unsystematic distortions of sampling error. In this section, we will review the theory of error of measurement and derive the classic formula for correction for attenuation. Measurement error has a special status among systematic artifacts: It is the only systematic artifact that is always present in every study in a meta-analysis.

We will then look at the impact of error of measurement on sampling error and confidence intervals. In particular, we will derive the confidence interval for individual corrected correlations. From this base, we will later consider the impact of error of measurement as it varies in amount from one study to another.

Let us denote by \( T \) the true score that would have been observed on the independent variable had we been able to measure it perfectly. We then have

\[
x = T + E_1
\]

where \( E_1 \) is the error of measurement in the independent variable. Let us denote by \( U \) the true score that would have been observed on the dependent variable had we been able to measure it perfectly. We then have

\[
y = U + E_2
\]

where \( E_2 \) is the error of measurement in the dependent variable. Let us use the traditional notation by denoting the reliabilities by \( r_{xx} \) and \( r_{yy} \), respectively. We then have

\[
\begin{align*}
r_{xx} &= \rho_{xT}^2 \\
r_{yy} &= \rho_{yU}^2
\end{align*}
\]
The desired correlation is the population correlation between perfectly measured variables, that is, $\rho_{TU}$, but the observed correlation is the sample correlation between observed scores $r_{xy}$. There are two steps in relating the one to the other: the systematic attenuation of the population correlation by error of measurement and the unsystematic variation produced by sampling error.

The systematic attenuation can be computed by considering the causal pathways from $x$ to $T$ to $U$ to $y$. According to the rules of path analysis, $\rho_{xy}$ is the product of the three paths from $x$ to $y$, that is,

$$\rho_{xy} = \rho_x \rho_T \rho_{TU} \rho_{TU} = \rho_x \rho_T \rho_{TU}$$

$$= \sqrt{r_{xx}} \sqrt{r_{yy}} \rho_{TU}$$

At the level of population correlations, this leads to the classic formula for correction for attenuation:

$$\rho_{TU} = \frac{\rho_{xy}}{\sqrt{r_{xx}} \sqrt{r_{yy}}} \quad (3.12)$$

At the level of observed correlations, we have

$$r_{xy} = \rho_{xy} + e$$

where $e$ is the sampling error in $r_{xy}$ as before. Thus,

$$\sigma_r^2 = \sigma_{\rho}^2 + \sigma_e^2$$

where the sampling error variance is given by the formulas of earlier sections.

If we correct the observed correlation using the population correlation formula, then we have

$$r_c = \frac{r_{xy}}{\sqrt{r_{xx}} \sqrt{r_{yy}}} = \frac{\sqrt{r_{xx} \sqrt{r_{yy}} \rho_{TU} + e}}{\sqrt{r_{xx} \sqrt{r_{yy}}}}$$

$$= \rho_{TU} + \frac{e}{\sqrt{r_{xx} \sqrt{r_{yy}}}} \quad (3.13)$$

We can write a new equation for the corrected correlation

$$r_c = \rho_c + e_c \quad (3.14)$$

where $e_c$ is the sampling error in the corrected correlation $r_c$ and where the population value $\rho_c = \rho_{TU}$. The error variance for the corrected correlation can then be computed from the error variance for uncorrected correlations and the reliabilities of the two variables:

$$e_c = \frac{e}{\sqrt{r_{xx} \sqrt{r_{yy}}}}$$

$$\sigma_{e_c}^2 = \frac{\sigma_e^2}{r_{xx} r_{yy}} \quad (3.15)$$

Thus, if we correct the observed correlation for attenuation, we increase the sampling error correspondingly. In particular, to form the confidence interval for a corrected correlation, we apply the correction formula to the two endpoints of the confidence interval for the uncorrected correlation. That is, in the case of correction for attenuation, just as we divide the point estimate of the correlation by the product of the square roots of the reliabilities, so, too, we divide each endpoint of the confidence interval by the same product.
An Example of Correction for Attenuation

Suppose that if organizational commitment and job satisfaction were perfectly measured, the correlation between true scores would be $\rho_{TU} = .60$. Suppose instead that we measure organizational commitment with reliability $r_{xx} = .45$ and that we measure job satisfaction with reliability $r_{yy} = .55$. Then the population correlation between observed scores would be

$$
\rho_{xy} = \sqrt{r_{xx}} \sqrt{r_{yy}} \rho_{TU} = \sqrt{.45 \times .55} \rho_{TU} = .50 \times 0.60 = .30
$$

That is, the effect of error of measurement in this example is to reduce the correlation between true scores by 50%—from a true score population correlation of .60 to a study population correlation of .30 between observed scores. If we apply the correction formula, we have

$$
\rho_{TU} = \frac{\rho_{xy}}{\sqrt{r_{xx}} \sqrt{r_{yy}}} = \frac{.30}{\sqrt{.45 \times .55}} = .30 \times .50 = .60
$$

That is, correction for attenuation works perfectly for population correlations; it is perfectly accurate when sample size is infinite.

Consider the impact of sampling error. If the sample size for the study is $N = 100$, then the standard error of the observed correlation (from $\rho_{xy} = .30$) is $(1 - .30^2)/\sqrt{99} = .091$. Thus, it would not be uncommon to observe a correlation of .20 in the actual study. If we compare the observed correlation of .20 with the desired correlation of .60, then we see that there is a massive error. However, this error can be broken into two components: the systematic error of attenuation and the unsystematic error due to sampling error. The systematic error reduced the correlation from .60 to .30. The unsystematic error is the difference between an observed $r$ of .20 and the population attenuated correlation .30.

Let us correct for attenuation and look at the error in the corrected correlation:

$$
\rho_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}} \sqrt{r_{yy}}} = \frac{.20}{\sqrt{.45 \times .55}} = .20 \times .50 = .40
$$

The sampling error in the corrected correlation is the difference between the estimated .40 and the actual .60. Thus, we have

$$
r = \rho_{xy} + e = \rho_{xy} - .1 = .30 - .10 = .20
$$

$$
r_e = \rho_e + e_e = \rho_e - .2 = .60 - .20 = .40
$$

Therefore, as we doubled the observed attenuated correlation to estimate the unattenuated correlation, we doubled the sampling error as well. On the other hand, we reduced the systematic error from .30 to 0. We can combine both types of error and look at total error. Total error has been reduced by 50%:

Total error for $r = .60 - .20 = .40$

Total error for $r_e = .60 - .40 = .20$
The standard error for the observed correlation would be calculated as

\[
\frac{1 - .20^2}{\sqrt{99}} = .096
\]

The 95% confidence interval for the observed correlation is given by \( r \pm 1.96\sigma_e = .20 \pm 1.96(.096) \) or \(.01 \leq \rho \leq .39\), which does include the actual value of \( \rho_{xy} = .30 \). We then correct each endpoint of the confidence interval to obtain

<table>
<thead>
<tr>
<th>Lower Endpoint</th>
<th>Upper Endpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_1 = .01 )</td>
<td>( r_1 = .39 )</td>
</tr>
<tr>
<td>( r_{1e} = \frac{.01}{\sqrt{.45} \sqrt{.55}} )</td>
<td>( r_{2e} = \frac{.39}{\sqrt{.45} \sqrt{.55}} )</td>
</tr>
<tr>
<td>( = .01 )</td>
<td>( = .39 )</td>
</tr>
<tr>
<td>( \sqrt{.50} )</td>
<td>( \sqrt{.50} )</td>
</tr>
<tr>
<td>( r_{1e} = .02 )</td>
<td>( r_{2e} = .78 )</td>
</tr>
</tbody>
</table>

Let us compare the confidence intervals of corrected and uncorrected correlations:

\[
.01 \leq \hat{\rho}_{xy} \leq .39
\]

\[
.02 \leq \hat{\rho}_{TU} \leq .78
\]

We see that the center of the confidence interval changes from the uncorrected correlation \(.20\) to the corrected correlation \(.40\). At the same time, the width of the confidence interval doubles, reflecting the increased sampling error in the corrected correlation.

This point can be made dramatically with confidence intervals. If measurement of both variables were perfectly reliable, then the population correlation would be \(.60\), and the standard error would be \((1 - .60^2)/\sqrt{99} = .064\). This is much smaller than the standard error for a correlation of \(.30\), which is \(.091\). Thus, if we can reduce error of measurement substantively, then we can obtain larger observed correlations and smaller confidence intervals. Substantive elimination of error of measurement is vastly superior to elimination by statistical formula after the fact.

We could have obtained the same confidence interval in a different way. Suppose we erected the confidence interval around the corrected correlation using the sampling error formula for \( e_c \). The center of the confidence interval is then \( r_c = .40 \). The sampling error variance is then given by

\[
\sigma_{e_c}^2 = \frac{\sigma_e^2}{r_{xx} r_{yy}} = \frac{(1 - .2^2)/99}{(.45)(.55)} = (1 - .2^2)/2475 = .0376
\]

That is, the sampling error standard deviation for the corrected correlation is \( \sigma_{ec} = (.0376)^{1/2} = .19\). The confidence interval is then given by \(.40 \pm 1.96\sigma_{ec} \) or \(.02 \leq \rho_{TU} \leq .78\). This is the same confidence interval obtained earlier.

**Statistical Versus Substantive Correction**

If we use statistical formulas to correct for attenuation, we obtain larger corrected correlations with a wider confidence interval. There are two conclusions that might
be drawn from this fact. (1) **False conclusion:** Because correcting for attenuation increases the amount of sampling error, maybe we should not correct for attenuation. **Key fact:** If we do not correct for attenuation, then we do not eliminate the systematic error. In our example, the error in the uncorrected correlation was $0.60 - 0.20 = 0.40$. Thus, the error in the corrected correlation was only half as large as the error in the uncorrected correlation. (2) **True conclusion:** We could greatly improve our statistical accuracy if we could reduce the error of measurement substantively, that is, by using better measurement procedures in the first place.

**Using the Appropriate Reliability Coefficient**

In making corrections for measurement error, the researcher should make every effort to use the appropriate type of reliability coefficient. Different methods of computing reliability coefficients assess different kinds (sources) of measurement error. Estimation of the appropriate reliability coefficient requires the specification of the kinds of measurement error in the specific research domain and requires the gathering of data that allows each kind of measurement error to be reflected in the reliability coefficient. In most current meta-analyses, the unit of analysis is the person (or rat or pigeon or . . .) and the variable measured is some behavior of the person. For this case, there is an extensive theory of reliability developed by psychometricians and there has been extensive testing of that theory (see, e.g., Cronbach, 1947; Stanley, 1971; Thorndike, 1951). This is the case discussed in this book. However, meta-analyses are being conducted in other areas. For example, Rodgers and Hunter (1986) conducted a meta-analysis in which the measures were business unit productivity measures. Another example of this is Harter, Schmidt, and Hayes (2002). In both meta-analyses, the unit of measurement was not the individual employee but the business unit (e.g., individual stores or individual bank branches), and in both meta-analyses special reliability estimation procedures had to be used. However, such cases are the exception. Suppose the unit of analysis in the study is persons, the usual case. The measure is usually obtained in one of three ways: the behavior is directly recorded (as in test scores or closed-ended questionnaire answers, here called “response data”), assessed by an observer (here called a “judgment or rating”), or observed and recorded by an observer (here called “coded response data”). The reliability considerations differ by case.

**Response Data.** Response data behavior is used for measurement of a given variable under the assumption that the variable to be measured is the primary causal agent that determines that behavior. Error of measurement is present to the extent that the behavior is determined by other causal agents. At least three kinds of error have been identified by psychometric theory: random response error, specific error, and transient error (cf. Le & Schmidt, 2002; Le, Schmidt, & Lauver, 2003; Schmidt & Hunter, 1996, 1999b; Schmidt, Le, & Ilies, 2003; Stanley, 1971; Thorndike, 1949, 1951). The ubiquitous error agent is randomness in behavior. Except for highly practiced responses, such as giving one’s name, most human acts have a sizeable random element. This is called “random response error.” Random response error can be thought of as noise in the human nervous system (Schmidt &
Hunter, 1999b; Thorndike, 1949, 1951). Sometimes the behavior is influenced by something peculiar about the measurement situation, for example, an idiosyncratic response to one of the specific words used in an opinion item. The influence of such situation- or stimulus-specific agents is called “specific error” (or specific factor error). Specific factor error is associated with individual items in a measure. (It is actually the person by item interaction.) Each different scale for measuring a variable also has specific factor measurement error unique to that scale. Sometimes the behavior is affected by an influence that varies randomly over time, such as mood or illness. The influence of a time-varying factor is called “transient error.” Each of these three forms of error of measurement enters differently into the designs traditionally used to measure reliability (Schmidt et al., 2003; Stanley, 1971; Thorndike, 1951).

One design makes error visible by obtaining multiple responses in a given measurement session—eliciting behaviors to various situations or stimuli (e.g., items) that are equivalent in the extent of causal influence from the variable to be measured. The independent responses allow for independent sampling of the random response error and independent sampling of the specific error. Thus, reliability computed from this design detects and measures the extent of random response error and specific error. However, if there is transient error, this design will not detect it; thus, the reliability coefficient will be too large by the relative amount of transient error. This form of reliability is properly called “parallel forms reliability” and is improperly called “internal consistency reliability,” although better language would substitute the phrase “reliability estimate” for the word “reliability.” This form of reliability is usually assessed using Cronbach’s alpha coefficient or the KR-20 coefficient (the equivalent coefficient for items answered dichotomously). Cronbach (1947) referred to this type of reliability as the coefficient of equivalence (CE).

Another common design is the test-retest design. A behavior is measured with the same scale at two points in time that are far enough apart so that the transient error factor is not repeated, but close enough in time so that there is no significant change in the variable to be measured. The two measures will allow for new (independent) sampling from the random response error distribution and new (independent) sampling from the transient error distribution. However, if there is specific error, this design will not detect it; thus, the reliability coefficient will be too large by the relative amount of specific error. Specific factor measurement error is not detected because the same specific errors occur at time 1 and time 2 due to use of the same scale (instrument) at both times. This form of reliability is called “test-retest reliability,” although better language would substitute the phrase “reliability estimate” for the word “reliability.” Cronbach (1947) referred to this type of reliability estimate as the coefficient of stability (CS).

If all three kinds of measurement error are present, then the correct reliability will be obtained only by a design in which both the situational determinants of specific error and the temporal determinants of transient error are resampled. This is the “delayed parallel forms” design. Cronbach (1947) referred to this type of reliability estimate as the coefficient of equivalence and stability (CES). Given two forms of the measure and two occasions, it is possible to separately estimate the extent of each of the three kinds of error (cf. Cronbach, 1947; Schmidt et al., 2003; Stanley, 1971; Thorndike, 1949, 1951). If all three kinds of error are present,
then the delayed parallel forms reliability (CES) will be smaller than either the parallel forms reliability estimate or the test-retest reliability estimate. The effect of using the wrong reliability estimate is to underestimate the impact of error of measurement. Thus, correction using the wrong reliability means that some source of error is not calibrated and, as a result, the correlation is not corrected for attenuation due to that error source. That is, correction using the wrong reliability means that the corrected correlation will correspondingly underestimate the actual correlation between constructs. This will, of course, cause a downward bias in estimates of mean correlations in the meta-analysis. If a primary researcher or a meta-analyst is forced to use CE or CS estimates of reliability because CES estimates are not available, the study or meta-analysis should point out that full correction for measurement error was not possible and that the results reported are therefore conservative (i.e., have a downward bias). The report should also point out that such an incomplete correction for measurement error yields results that are much more accurate than failure to make any correction for measurement error.

Judgments by Raters. If a construct such as job performance is assessed by an observer such as the person’s immediate supervisor, then there are two sources of error in the measurement: error in the judgment and idiosyncrasy in rater perception (called “halo”; cf. Hoyt, 2000; Schmidt, Viswesvaran, & Ones, 2000; Viswesvaran, Ones, & Schmidt, 1996; Viswesvaran, Schmidt, & Ones, in press). First, the judgment is itself a response by the rater and is, thus, subject to random response error and potentially subject to specific error in rating scale items and to transient error. These latter sources are the only sources of error assessed by data methods that consider only data from one rater. Second, in no research area has human perception of other people proved to be without idiosyncrasy. In fact, in most areas, there are large differences in the perceptions of different raters. This idiosyncratic bias associated with each rater (halo) is not part of the construct being assessed and, thus, is measurement error (Hoyt, 2000; Viswesvaran et al., 1996, in press). This form of measurement error functions as specific factor error associated with each rater. It is also referred to as halo error. Thus, the proper reliability estimate must take differences in perception into account as well as randomness in the judgment made by each judge. The appropriate way to estimate reliability for judgments is to correlate judgments made independently by different raters. The correlation between judges will be reduced to the extent that there is either random measurement error in the judgments or differences in perception (halo error). If two judgments (say, “quality of work” and “quantity of work”) are each made by two raters, then it is possible to separately estimate the effect of random response error, item-specific error (specific to these two rating scale items), and idiosyncrasy error (also called “halo”; Viswesvaran et al., 1996, in press). If there is transient error, and if the cause of the transient error is independent in the two judges (the usual case), then this correlation will be appropriately lowered by transient error also, although transient error will not be distinguishable from idiosyncrasy (halo) error. In this case, the reliability is called “inter-rater reliability,” although it, too, would better be called a reliability estimate. Inter-rater reliability is the CES (delayed parallel forms reliability) for ratings. Studies of this sort have found average inter-rater reliabilities of about .50 for supervisory ratings of job performance (cf. Rothstein, 1990; Schmidt et al., 2000; Viswesvaran
et al., 1996). Intra-rater reliability, estimated by correlating ratings made by the same rater at two different times, is much higher (around .85; Viswesvaran et al., 1996) but is a gross overestimate of actual reliability.

**Coded Response Data.** Some behaviors are too complicated to be directly recorded. Thus, the data used represent a coding by an observer of the behavior. For example, we might code the extent of need for achievement expressed in a story inspired by a picture shown to the subject. Differences between the codings by different observers are called “coding error.” A critical error often made in this context is to consider as error only the discrepancy between coders. While coding error is one important source of error in the measurement, there are also random response error, specific error, and transient error in the behavior coded. For example, suppose that coders were so well trained that they agreed to a correlation of .95 in their assessments of the amount of achievement motivation contained in the stories. However, suppose that from one week to the next, there is only a correlation of .40 between the achievement imagery in successive stories told by the same person. The “reliability” of .95 would not reflect this instability (random response error and transient error) in inventing stories and would, thus, greatly overestimate the actual reliability in question. For a more complete treatment of measurement error in coded response data, see Schmidt and Hunter (1996, 1999b).

What are the implications of these facts about measurement errors and reliability estimates for meta-analysis? Unlike other systematic artifacts, measurement error is always present. Accurate corrections for measurement error are critical to accurate meta-analysis results (Cook et al., 1992, pp. 215–216). The use of inappropriate reliability estimates to make corrections for measurement error in meta-analysis leads to some loss of accuracy of results. For example, recent research indicates that there is transient measurement error in commonly used psychological measures (e.g., personality and ability measures) (Schmidt et al., 2003). This means that the common practice of using coefficient alpha and KR-20 reliability estimates to correct for measurement error in meta-analysis typically leads to undercorrections and, hence, produces a downward bias in estimates of mean correlations, because coefficient alpha and KR-20 reliabilities do not detect or remove the effects of transient error. Likewise, use of test-retest reliability estimates will also lead to undercorrections, because of failure to control for specific factor measurement errors. Whether one is correcting each coefficient individually (as described in this chapter) or using distributions of reliability coefficients (as described in Chapter 4), all relevant sources of measurement error should be considered and, to the extent possible given available data, appropriate reliability estimates should be used (Schmidt et al., 2003).

Unfortunately, most available estimates of reliability are alpha coefficients or KR-20 estimates (i.e., estimates of the CE) and so do not take transient error into account. Estimates of the CES coefficient are rare in the literature and not common even in test manuals. Estimates of the amount of transient error in a variety of scales are around 4% to 5% (see, e.g., Schmidt et al., 2003). (These values are larger for trait measures of affectivity.) Hence, CE estimates can be adjusted or corrected for transient error by subtracting .04 or .05 from these reliability estimates, as
Meta-Analysis of Correlations Corrected Individually for Artifacts

suggested in Schmidt et al. (2003). Such adjustment figures, however, are not (yet) available for all types of scales. Therefore, it is important to bear in mind that in research, as in other areas, “the perfect is the enemy of the good.” It would be a false argument to state that unless ideal estimates of reliability can be used, no correction for measurement error should be made. As noted earlier, use of coefficient alpha estimates (CE estimates) when CES estimates (delayed parallel forms estimates) are more appropriate still leads to final results that are much more accurate than failure to make any correction for the biasing effects of measurement error. However, in such a case, the researcher should point out that only a partial correction for measurement error was possible and the resulting corrected values therefore contain a downward bias.

Another important implication is that when ratings are used, use of intra-rater reliabilities will lead to very severe undercorrections for measurement error. To avoid this, inter-rater reliabilities should always be used, and (except in very special cases) intra-rater reliabilities should not be used. Also, in artifact distribution meta-analysis (discussed in Chapter 4), one should never use a mixture of intra- and inter-rater reliabilities in the reliability artifact distribution. Fortunately, inter-rater estimates (unlike CES estimates for tests and scales) are widely available (cf. Rothstein, 1990; Viswesvaran et al., 1996).

In the case of coded response data, it is critical that the reliability estimates be based on two separate administrations of the measure. That is, the first coder should code responses obtained at time 1 and the second coder should code responses from the same individuals obtained at time 2. The correlation between these two codings provides the only accurate estimate of the reliability of coded response data. If both coders code responses obtained on a single occasion, the resulting correlation grossly overestimates reliability, resulting in a large downward bias in meta-analysis estimates.

The use of inappropriate reliability estimates has been a problem in some applications of meta-analysis. It is important to remember that different types of reliability estimates are not all equally appropriate.

Restriction or Enhancement of Range

If studies differ greatly in the range of values present on the independent variable, then the correlation will differ correspondingly. Correlations are directly comparable across studies only if they are computed on samples from populations with the same standard deviation on the independent variable. Range correction formulas are available that take a correlation computed on a population with a given standard deviation and produce an estimate of what the correlation would have been had the standard deviation been different. That is, range correction formulas estimate the effect of changing the study population standard deviation from one value to another. To eliminate range variation from a meta-analysis, we can use range correction formulas to project all correlations to the same reference standard deviation.

As noted in Chapter 2, range restriction (or range enhancement) can be either direct or indirect—and range correction procedures are different for the two cases.
Direct range restriction occurs when there is direct truncation on the independent variable. For example, if only those in the top 50% of test scores are hired, and no one with a lower test score is hired, we have direct range restriction. Likewise, it is direct range enhancement if an experimenter selects as subjects only those in the top 10% and the bottom 10% of scores on, say, a measure of conscientiousness to participate in a study. The statistic we need to know to correct for direct range restriction is the ratio of the observed SD in the restricted sample to the observed SD in the unrestricted sample, that is, \( s_x / S_X \). This ratio is referred to as \( u_X \); that is, \( u_X = s_x / S_X \). If we know this ratio, we can use the Thorndike Case II (Thorndike, 1949) direct range restriction formula to correct for range restriction. By using the reciprocal of this statistic, that is, \( S_X / s_x \), we can use this same formula to correct for direct range enhancement. This ratio is referred to as \( U_X \); that is, \( U_X = S_X / s_x \).

Indirect range restriction occurs when people are selected on a third variable that correlates with the independent variable. For example, if we are evaluating the validity of a new engineering aptitude test in an engineering college and all our engineering students were originally admitted based on a college entrance examination, then there will be indirect range restriction on the engineering aptitude test. If selection into the college is based only on the entrance examination (direct range restriction on that exam), and if we know the restricted and unrestricted SD on that examination, there is a formula for correcting for this sort of indirect range restriction (called the Thorndike Case III indirect range restriction correction formula; Thorndike, 1949). However, this situation is very rare, because selection on the third variable (here the college entrance exam) is rarely direct (because other variables are also used in selection) and because, even if it is, we rarely have the needed statistics on the third variable. Thus, this correction formula can rarely be applied. For this reason, we do not develop applications of this formula.

The type of indirect range restriction that is most common is the situation in which people have been selected on an unknown (unrecorded) combination of variables and that combination (or composite) of variables is correlated with the independent variable, producing indirect range restriction on the independent variable (Linn, Harnisch, & Dunbar, 1981a). A common example is the case in which people have been hired based on some unknown combination of, say, an interview, an application blank, and a background check, and we find that the SD on the test we are studying is considerably smaller in this selected (incumbent) group than the SD in an application population (the unrestricted SD), indicating there is indirect range restriction on our independent variable. That is, we find \( u_x \) is less than 1.00. Linn et al. (1981a) showed that students admitted to law schools were an example of this sort of indirect range restriction. Another example is the case in which people who volunteer to participate in a study have a lower SD on extroversion scores (the independent variable); that is, \( u_x \) is again less than 1.00. Here, indirect range restriction is produced by some unknown combination of self-selection variables. Most range restriction in real data is, in fact, caused by this sort of indirect range restriction (Hunter, Schmidt, & Le, 2002; Linn et al., 1981a; Mendoza & Mumford, 1987; Thorndike, 1949). As in the case of direct selection, we can also have indirect range enhancement. For example, the SD on extroversion of those volunteering for the study might be larger than the SD of
extroversion in the reference population. However, in the case of indirect selection, range enhancement appears to be relatively rare.

The key statistic we must know to correct for indirect range restriction of this sort is not $u_X$ but $u_T$, where $u_T$ is the ratio of true score SDs. That is, $u_T = s_T / S_T$. As we will see, there are special formulas for computing $u_T$ from $u_X$ and other information. If we have indirect range enhancement (instead of range restriction), the statistic we need is $U_T = S_T / s_T$. The correction for indirect range restriction is made using the same correction formula used to correct for direct range restriction, but with $u_T$ used in that formula in place of $u_X$. Likewise, to correct for range enhancement, $U_T$ is substituted in that formula for $U_X$. Procedures for correcting for this type of range restriction—the most commonly occurring type of range restriction—have been developed only recently (Hunter et al., 2002; Mendoza & Mumford, 1987) and were not included in the 1990 edition of this book. The availability of these corrections greatly increases the accuracy of meta-analyses involving range restriction corrections. More detailed information on indirect range restriction can be found in Chapter 5.

For each study, we need to know the standard deviation of the independent variable $s_x$. Range departure is then measured by relating that standard deviation to the reference standard deviation $S_x$. The comparison used is the ratio of the standard deviation in the study group to the reference standard deviation, that is, $u_X = s_x / S_x$. The ratio $u_X$ is less than 1 if the study has restriction in range (the usual case) and greater than 1 if the study has enhancement of range. The correlation in the study will be greater than or less than the reference correlation depending on whether the ratio $u_X$ is greater than or less than 1, respectively.

In direct range restriction, the correction for range restriction depends on the value of $u_X$. In indirect range restriction, the correction depends on $u_T$, which is a function of $u_X$ and $r_{xxa}$, the reliability of the independent variable in the unrestricted population. These corrections depend on two assumptions. First, the relationship in question must be linear (or at least approximately so). Second, the variance of the independent variable must be equal (or at least approximately so) at each level of the dependent variable. This latter condition is known as homoscedasticity (Gross & McGanney, 1987). As we will see in Chapter 5, one additional assumption is required for the correction for indirect range restriction.

As noted previously, the major computational difference between direct and indirect range restriction is that in direct range restriction we use $u_X$ in making range corrections and in indirect range restriction we use $u_T$. The correction formula is otherwise identical. (There are also some differences, described in this chapter, in the order in which corrections for measurement error are made.)

The extra steps involved in computing $u_T$ make the mathematics of correcting for indirect range restriction somewhat more complicated than is the case for direct range restriction and, therefore, make the presentation somewhat more mathematically complicated and lengthy. However, once $u_T$ is computed and substituted for $u_X$, the correction equation is the same. Likewise, the computation of sampling error variance and confidence intervals for range-corrected correlations is also identical in form. Therefore, to simplify our presentation, we will present most of the following discussion in terms of direct range restriction corrections and will forgo the presentations for indirect range restriction corrections at this point.
META-ANALYSIS OF CORRELATIONS

(We will return to the topic of indirect range restriction later in this chapter.) Before proceeding, we present the necessary formulas for computing \( u_T \):

\[
\begin{align*}
\mathit{u}_T &= \frac{s_T}{S_T} \\
\mathit{u}_T &= \left[ \frac{u_X^2 - (1 - r_{XX_a})}{r_{XX_a}} \right]^{1/2} \quad (3.16)
\end{align*}
\]

where \( r_{XX_a} \) is the reliability of the independent variable in the unrestricted group. If only the reliability in the restricted group \( (r_{XX_i}) \) is known, \( r_{XX_a} \) can be computed as follows:

\[
\begin{align*}
\mathit{r}_{XX_a} &= 1 - \frac{s_{XX_i}^2 (1 - r_{XX_i})}{s_{XX_a}^2} \quad (3.17a) \\
\mathit{r}_{XX_a} &= 1 - \frac{u_X^2 (1 - r_{XX_i})}{UX^2} \quad (3.17b)
\end{align*}
\]

This equation can be reversed to give \( r_{XX_i} \) for any value of \( r_{XX_a} \):

\[
\mathit{r}_{XX_i} = 1 - \frac{U_X^2 (1 - r_{XX_a})}{UX^2} \quad (3.17c)
\]

where \( U_X = 1/u_X \). (Equation [3.16] is derived in Chapter 5, where it appears as Equation [5.31].)

In this section, we will use the direct range restriction model to explore range departure in the context of a single study in which population correlations are known. In the following section, we will consider the effect of correcting a sample correlation for range departure. We will find that the sampling error of the corrected correlation differs from that of the uncorrected correlation, and we will show how to adjust the confidence interval correspondingly. We will then briefly note the relationship between the ratio \( u_X \) and the “selection ratio” in personnel selection research. After this treatment of range correction in single studies, we will consider the effect of range correction in meta-analysis. At that point, we will treat both indirect and direct range restriction.

We cannot always study the population that we wish to use as a reference point. Sometimes we study a population in which our independent variable varies less than in the reference population (restriction in range) and sometimes we study a population in which it varies more widely than in the reference population (enhancement of range). In either case, the same relationship between the variables produces a different correlation coefficient. In the case of enhancement of range, the study population correlation is systematically larger than the reference population correlation. This problem is compounded by sampling error and by error of measurement.

Consider personnel selection research. The reference population is the applicant population, but the study is done with people who have already been hired (because we can get job performance scores only for those on the job). If the people hired were a random sample of the applicants, then the only problems would be sampling error and measurement error. Suppose, however, the test we are studying has been used to select those who are hired. For example, suppose those hired are those who are above the mean on the test; that is, we have direct range restriction. Then the range of test scores among the job incumbents is greatly reduced in comparison to
the applicant population. We would thus expect a considerable reduction in the size of the population correlation in the incumbent population compared to the applicant population. If test scores are normally distributed in the applicant population, then the standard deviation for people in the top half of the distribution is only 60% as large as the standard deviation for the entire population. Thus, if the standard deviation were 20 in the applicant population, it would be only $.60(20) = 12$ in the incumbent population of those hired. The degree of restriction in range would thus be $u_X = 12/20 = .60$.

The formula for the correlation produced by a direct selection change in distribution in the independent variable is called the formula for restriction in range, although it works for enhancement, too, as we shall see. Let $\rho_1$ be the reference population correlation and let $\rho_2$ be the study population correlation. Then

$$\rho_2 = \frac{u_X \rho_1}{\sqrt{(u_X^2 - 1)\rho_1^2 + 1}}$$

(3.18)

where

$$u_X = \frac{\sigma_2}{\sigma_1}$$

is the ratio of standard deviations in the two populations. In the case of direct restriction in range, we have $u_X < 1$ and, hence, $\rho_1 > \rho_2$. In the case of direct range enhancement, we have $u_X > 1$ and, hence, $\rho_2 > \rho_1$.

In the case of our personnel selection example, we have $u_X = .60$ and, hence,

$$\rho_2 = \frac{.60\rho_1}{\sqrt{(.60^2 - 1)\rho_1^2 + 1}} = \frac{.60\rho_1}{\sqrt{1 - .64\rho_1^2}}$$

For example, if the correlation between test and job performance in the applicant population were .50, then the correlation in the study population would be

$$\rho_2 = \frac{.60(.50)}{\sqrt{1 - .64(.50)^2}} = \frac{.60(.50)}{.92} = .33$$

That is, if the study were done on only the top half of the distribution on the independent variable, then the population correlation would be reduced from .50 to .33. If undetected, this difference between .50 and .33 would have profound implications for the interpretation of empirical studies.

Suppose, however, we have the data, that is, $\rho_2 = .33$ and $u_X = .60$, and we wish to correct for restriction in range. We could reverse the roles of the two populations. That is, we could regard the applicant population as an enhancement of the incumbent population. We could then use the same formula as before with the $\rho$s reversed, that is,

$$\rho_1 = \frac{U_X \rho_2}{\sqrt{(U_X^2 - 1)\rho_2^2 + 1}}$$

(3.19)

where

$$U_X = \frac{\sigma_1}{\sigma_2} = \frac{1}{u_X}$$
is the ratio of standard deviations in the opposite order. This formula is called the correction for restriction in range, although it also works for correction for enhancement. In the personnel example, we plug in \( \rho_2 = .33 \) and \( U = 1/u_X = 1/.60 = 1.67 \) to obtain

\[
\rho_1 = \frac{1.67(.33)}{\sqrt{(1.67^2 - 1)(.33)^2 + 1}} = \frac{1.67(.33)}{1.09} = .50
\]

Thus, at the level of population correlations (i.e., when \( N \) is infinite), we can use the formula for restriction in range to move back and forth between populations of different variance with perfect accuracy. If range restriction were indirect, substituting \( u_T \) and \( U_T \) for \( u_X \) and \( U_X \), respectively, would allow us to do this same thing.

The situation is more complicated if there is sampling error. If we apply the formula for correction for restriction in range to a sample correlation, then we get only an approximate estimate of the reference group population correlation. Moreover, the corrected correlation will have a different amount of sampling error. This situation is analogous to that in correction for attenuation due to measurement error. There is a trade-off. In order to eliminate the systematic error associated with restriction in range, we must accept the increase in sampling error produced by the statistical correction. If we could correct substantively, that is, if the study could be done on the reference population, then there would be no increase in sampling error. In fact, in the case of restriction in range, the study done on the applicant population (if it could be done) would have the larger correlation and, hence, the smaller confidence interval.

The confidence interval for the corrected correlation is easy to obtain. The correction formula can be regarded as a mathematical transformation. This transformation is monotone (but not linear) and, hence, it transforms confidence intervals. Thus, the confidence interval is obtained by correcting the endpoints of the confidence interval using the same formula that is used to correct the correlation. That is, the same range correction formula that is applied to the correlation is applied to the endpoints of the confidence interval.

Consider the personnel example with direct range restriction in which the population correlations are .50 for the applicant population and .33 for the study population. If the sample size is 100, then the standard error for a correlation of \( \rho = .33 \) is \( \sigma_c = (1 - .33^2)/\sqrt{99} = .09 \). If the sample correlation came out low, it might be something such as .28, which is low by .05. Corrected for restriction in range of \( U_X = 1.67 \), we have

\[
r_c = \frac{1.67(.28)}{\sqrt{(1.67^2 - 1)(.28^2) + 1}} = .47 = .44
\]

The standard error of the observed \( r \) of .28 is .093. The 95% confidence interval on the observed \( r \) of .28 is

<table>
<thead>
<tr>
<th>Lower Endpoint</th>
<th>Upper Endpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_1 = .10 )</td>
<td>( r_2 = .46 )</td>
</tr>
<tr>
<td>( r_{c1} = \frac{1.67(.10)}{\sqrt{(1.67^2 - 1).10^2 + 1}} = .16 )</td>
<td>( r_{c2} = \frac{1.67(.46)}{(1.67^2 - 1).46^2 + 1} = .65 )</td>
</tr>
</tbody>
</table>
Thus, the confidence interval for the corrected correlation is $0.16 \leq \rho_c \leq 0.65$, which includes the actual value of $\rho_c = 0.50$. This confidence interval is much wider than the confidence interval for the uncorrected correlation and wider yet than the confidence interval that would have been found had the study been done in the reference population itself.

In this example, the range restriction is direct. We remind the reader that if the range restriction were indirect, one would use $u_T$ in place of $u_X$. Otherwise, the procedures would be identical.

**Range Correction and Sampling Error**

There is no difficulty in obtaining a confidence interval for a corrected correlation using the range correction formula; we simply correct the two endpoints of the confidence interval for the uncorrected correlation. However, it is not so easy to compute the standard error of the corrected correlation. Correction for attenuation due to measurement error is a linear operation; the uncorrected correlation is just multiplied by a constant. Thus, the sampling error and the error standard deviation are multiplied by the same constant. However, the range correction formula is not linear, and there is no exact formula for the resulting standard error. (The nature of the nonlinearity is that, for the same value of $u_X$, the correction increases smaller correlations by a greater percentage increase than it increases larger correlations.) The extent of nonlinearity depends on the size of the numbers involved, that is, the extent to which $U_X$ is different from 1 and the extent to which the uncorrected correlation has a square much greater than 0. If the nonlinearity is not too great, then we can approximate the sampling error by pretending that we have just multiplied the uncorrected correlation by the constant $\alpha = \frac{r_c}{r}$.

The sampling error would then be approximately

$$\sigma^2_{\rho_c} = \alpha^2 \sigma^2_{\rho}$$

(3.20)

To see the extent of this approximation, let us consider our personnel research example. We center our confidence interval for the corrected correlation about the corrected correlation itself, that is, around $r_c = 0.44$. The error standard deviation for the uncorrected correlation is $(1 - 0.28^2)/\sqrt{99} = 0.093$, and the ratio of corrected to uncorrected correlations is $0.44/0.28 = 1.57$. Hence, the estimated standard error for the corrected correlation is $(1.57)(0.093) = 0.146$. The corresponding confidence interval is $0.15 \leq \rho_c \leq 0.73$. This implied confidence interval differs only slightly from the confidence interval obtained by correcting the endpoints, that is, $0.16 \leq \rho_c \leq 0.65$.

There is a more accurate estimate of the standard error that can be obtained using Taylor’s series as suggested by Raju and Brand (2003) and Raju, Burke, and Normand (1983). For large sample size, the sampling error in the corrected correlation induced by the sampling error in the uncorrected correlation is proportional to the derivative of the correction function. Whereas the correlation is multiplied by the constant $\alpha$, the standard deviation is multiplied by the number $aa$, where

$$a = 1/[(U_X^2 - 1)r^2 + 1]$$

(3.21)
The variance would be multiplied by $a^2\alpha^2$. In the personnel selection example, we have $\alpha = 1.57$ and

$$a = 1/[(1.67^2 - 1)(.28)^2 + 1] = 1/1.0352 = .8770$$

Thus, the standard error is multiplied by $.8770(1.57) = 1.38$ instead of 1.57. The standard error of the corrected correlation is therefore estimated as $(.8870)(1.57)(0.93) = .130$. The confidence interval found using the improved estimate of the standard error is

$$-.19 < \rho < .69$$

This is in comparison to the correct interval obtained by correcting the confidence interval endpoints, which was

$$-.16 < \rho < .65$$

This improved estimate of the standard deviation is barely worth the trouble for hand calculations, although it is easy to introduce into computer programs, and we have done so. The Windows-based meta-analysis program VG6 (see the Appendix for a description) contains this refinement. Again, we note that if the range restriction were indirect, $U_T$ would be used in the preceding formula for $a$ in place of $U_X$. (Bobko & Reick, 1980, also presented an equation for the standard error of a correlation corrected for range restriction; their formula appears different on the surface but produces the same results as our procedure. The same is true for the formulas presented by Cureton, 1936; Kelly, 1947; and Raju & Brand, 2003. Mendoza, Stafford, & Stauffer, 2000, presented a method of estimating the confidence interval for a corrected correlation without estimating its $SE$. See also Forsyth & Feldt, 1969.)

An Example: Confidence Intervals. Consider a personnel selection validation study with direct range restriction and using job performance ratings by a single supervisor. Given an observed correlation of .30 with a sample size of 100, the confidence interval for the uncorrected validity coefficient is $P [.12 \leq \rho \leq .48] = .95$. From King, Hunter, and Schmidt (1980), we know that the reliability of the supervisor ratings in the applicant pool is at most .60. If the selection ratio is 50%, then the formulas in Schmidt, Hunter, and Urry (1976) (presented later in this chapter) show that the ratio of the standard deviation of the applicant group to that of the incumbent population ($U_X$) is 1.67. The point correction of the observed validity coefficient is therefore

$$r_1 = \frac{1.67r}{\sqrt{(1.67^2 - 1)r^2 + 1}} = .46$$

$$r_2 = \frac{r_1}{\sqrt{.60}} = .60$$

The confidence interval for the corrected validity is obtained by applying the same corrections to the endpoints of the confidence interval for the uncorrected validity:
Lower Endpoint | Upper Endpoint
---|---
\( r_1 = \frac{1.67(1.2)}{\sqrt{(1.67^2 - 1).12^2 + 1}} \) | \( r_1 = \frac{1.67(4.8)}{\sqrt{(1.67^2 - 1).48^2 + 1}} \)
\( = .20 \) | \( = .67 \)
\( r_2 = \frac{.20}{\sqrt{.60}} = .26 \) | \( r_2 = \frac{.67}{\sqrt{.60}} = .86 \)

Hence, the confidence interval for the corrected validity is

\[
P\{.26 \leq \rho \leq .86\} = .95
\]

Direct Range Restriction and the Selection Ratio. In personnel research, restriction in range sometimes comes about in a very particular way: People are hired from the top down using the test that is to be validated. Only those hired appear in the validation study. Thus, those who appear in the study are chosen from the top portion of the reference population distribution of test scores (direct range restriction). Because the test score distribution in applicant populations is typically a normal or nearly normal distribution, the range restriction parameter \( u_X \) can be computed indirectly from the selection ratio.

The selection ratio is defined as the proportion of applicants selected by the test. For example, if all applicants in the top tenth of the distribution are offered employment, then the selection ratio is 10%. The test selection ratio will be equal to the percentage of applicants offered jobs if hiring is based solely on test scores from the top down (Schmidt, Hunter, McKenzie, & Muldrow, 1979).

Let \( p \) be the selection ratio as a proportion (i.e., as a fraction such as .10, rather than as a percentage). If we are hiring from the top of a normal distribution, then, corresponding to any selection ratio \( p \), there is a cutoff score \( C \) such that

\[
P[x \geq C] = p \tag{3.22}
\]

If that cutoff score is given in standard score form, then it can be looked up using any normal distribution table. Once the cutoff score is known, then we can compute the mean and variance in test scores among those selected in standard score form using the following formulas:

\[
\mu_x = \frac{\phi(C)}{p} \tag{3.23}
\]

where \( \phi(C) \) is the value of the unit normal density function at the cutoff (also called the “normal ordinate”) and

\[
\sigma_x^2 = 1 - \mu_x (\mu_x - C) = 1 - \mu_x^2 + C \mu_x \tag{3.24}
\]

Because the applicant population has a variance of 1 in standard scores, the number \( \sigma_x \) is equal to the parameter \( u \) in the range restriction formula.

For example, if the selection ratio is 10%, then the normal distribution table shows that a cutoff score of 1.28 is required to select the top tenth. The mean standard score among those selected will be

\[
\mu_x = \frac{1}{p} \sqrt{\frac{1}{2\pi}} e^{-C^2/2} = \frac{10}{2.507} e^{-0.82} = 1.76
\]
The variance among those selected is

\[ \sigma^2_x = 1 - 1.76^2 + 1.28(1.76) = .1552 \]

The standard deviation—and, hence, the parameter \( u_X \)—is then the square root of .1552, which is .39. That is, with a selection ratio of 10%, the standard deviation in the study population will be only 39% as large as the standard deviation in the applicant population.

It is important to note that the procedures described in this particular section apply only to direct range restriction. If range restriction is indirect, estimates of \( u_X \) will be somewhat inaccurate.

**Dichotomization of Independent and Dependent Variables**

The mathematics of dichotomization is very similar to that of correction for attenuation and will thus be developed succinctly. Some aspects of dichotomization were discussed in Chapter 2; a more detailed treatment is presented in Hunter and Schmidt (1990b) and MacCallum, Zhang, Preacher, and Rucker (2002). The key fact is that the impact of dichotomizing a continuous variable is to multiply the population correlation by an attenuating factor. This systematic attenuation can be corrected by dividing the attenuated correlation by the same factor. That is, if we know the factor by which the study correlation was attenuated, then we can restore the study correlation to its original value by dividing by that same attenuation factor.

If we divide a variable by a constant, then the mean and the standard deviation are divided by that same constant. Thus, the corrected correlation coefficient has a mean that is divided by the attenuation factor and a sampling error that is divided by the attenuation factor. Thus, the sampling error in the corrected correlation is larger than the sampling error in the uncorrected correlation. However, there is no other way to eliminate the systematic error introduced by the dichotomization.

Consider an example. Suppose the independent variable is split at the median. Then the attenuation factor is .80, and thus, the population correlation is reduced by 20%. If \( \rho \) is the true population correlation and \( \rho_o \) is the attenuated population correlation, then

\[ \rho_o = .80 \rho \]

This equation is algebraically reversible. To undo multiplication by .80, we divide by .80:

\[ \rho_o / .80 = (.80 \rho) / .80 = \rho \]

That is,

\[ \rho = \rho_o / .80 \]

is the formula for correction for dichotomization. The formula works perfectly for population correlations and works to eliminate the systematic error in the sample correlation.
The study sample correlation $r_o$ is related to the study population correlation in the usual manner:

$$r_o = \rho_o + e_o$$

where $e_o$ is the usual sampling error. If we correct the point biserial correlation to eliminate the attenuation due to dichotomization, then the corrected correlation $r_C$ is given by

$$r_C = r_o/.80 = (\rho_o + e_o)/.80 = \rho_o/.80 + e_o/.80$$

$$= (.80\rho)/.80 + e_o/.80$$

$$= \rho + e_o/.80$$

Let us denote by $e_C$ the sampling error in the corrected correlation. We then have

$$r_C = \rho + e_C$$

Thus, the population correlation corresponding to the corrected sample correlation is the desired true correlation; that is, the systematic part of the sample correlation is restored to its pre-dichotomization value $\rho$. However, the sampling error $e_C$ is not the usual sampling error associated with a population correlation of $\rho$. Rather, $e_C$ is the sampling error associated with a corrected correlation. Had there been no dichotomization to begin with, the standard deviation of the sampling error (the standard error) would have been

$$\sigma_e = (1 - \rho^2)/\sqrt{(N - 1)}$$

Instead, the standard deviation of the sampling error in the corrected correlation $\sigma_{eC}$ must be computed from the sampling error of the uncorrected correlation $\sigma_{eo}$. The sampling error standard deviation of the uncorrected correlation is

$$\sigma_{eo} = (1 - \rho_o^2)/\sqrt{(N - 1)}$$

$$= [1 - (.80\rho)^2]/\sqrt{(N - 1)}$$

$$= [1 - .64\rho^2]/\sqrt{(N - 1)}$$

The sampling error standard deviation of the corrected correlation is

$$\sigma_{eC} = \sigma_{eo}/.80 = 1.25\sigma_{eo}$$

Consider an example. Suppose the population correlation for the original continuous variables is $\rho = .50$. The population correlation with the independent variable split at the median is

$$\rho_o = .80\rho = .80(.50) = .40$$

If the sample size is $N = 100$, then the sampling error standard deviation for non-dichotomized variables is

$$\sigma_e = (1 - .50^2)/\sqrt{99} = .0754$$
The sampling error standard deviation of the uncorrected correlation is

$$\sigma_{eo} = (1 - .40^2) / \sqrt{99} = .0844$$

The sampling error standard deviation of the corrected correlation is

$$\sigma_{eC} = \sigma_{eo} / .80 = .0844 / .80 = .1055.$$  

Whereas 95% of sample correlations for non-dichotomized variables will be spread over

$$0.35 < r < 0.65$$

the corrected correlations spread over the range

$$0.29 < r_C < 0.71$$

For a more extreme split, the cost of correction will be higher. For a 90–10 split, the attenuation factor is .59, and the contrasting probability intervals are much different from each other:

$$0.35 < r < 0.65 \text{ if no dichotomization}$$
$$0.20 < r_C < 0.80 \text{ if one variable is split 90–10}$$

The situation is even more extreme if both variables are dichotomized. Consider a case from personnel selection. Hunter and Hunter (1984) found an average correlation of .26 between (continuous) reference recommendations and job performance ratings. Suppose (hypothetically) that there were no errors of measurement in the study and that, for purposes of communicating to the employer, the company psychologist decided to dichotomize the two variables. He dichotomizes the reference variable into “generally positive” versus “generally negative.” He finds that 90% of past employers give positive ratings while 10% give negative ratings. He splits the supervisor performance ratings at the median to produce “above average” versus “below average.” The effect of the double dichotomization (Hunter & Schmidt, 1990b) is to attenuate the correlation of .26 to

$$\rho_o = (.59)(.80)\rho = .472\rho = (.472)(.26) = .12$$

The corrected correlation is thus

$$r_C = r_o / .472 = 2.12 r_o$$

That is, the observed correlation must be more than doubled to correct the attenuation produced by the dichotomization. The sampling error is correspondingly increased. For a sample size of $N = 100$, the 95% confidence intervals in sample correlations are $0.08 < r < 0.44$ if the variables are not dichotomized and $-.15 < r_C < .67$ for a 10–90 and a 50–50 split.

Furthermore, it is not likely that the reference evaluations in a local validation study will have reliability as high as that in the Hunter and Hunter (1984) data. The reference check studies reviewed by Hunter and Hunter checked across three or
more past employers and used professionally developed scales to assess employer
ingutes. The correlation of .26 is corrected for the attenuation due to error of
measurement in the performance ratings only. Suppose the validation study asks
for references from just one past employer and uses performance ratings by the
immediate supervisor. If good rating scales are used in both cases, the reliability
of each variable would be expected to be at most .60. Because both reliabilities
are equal, the square roots are also equal at .77. The attenuation factor for error of
measurement is thus

\[ \rho_o = a_1a_2 \rho = (.77)(.77)\rho = .60\rho = (.60)(.26) = .156 \]

This correlation is then attenuated by double dichotomization

\[ \rho_{oo} = a_3a_4 \rho_o = (.59)(.80)\rho_o = .472\rho_o = .472(.156) = .074 \]

The net attenuation for both error of measurement and dichotomization is

\[ \rho_{oo} = (.60)(.472)\rho = .2832\rho \]

This value of .2832 is thus the attenuation factor for the correction of the sample
correlation. That is, the corrected correlation \( r \) will be

\[ r_C = r_{oo}/.2832 = 3.53 \]

That is, one must more than triple the observed sample correlation to correct for the
attenuation produced by both error of measurement and double dichotomization.
The sampling error is similarly increased. For a sample size of \( N = 100 \), the 95%
probability intervals are

\[ .08 < r < .44 \] if perfect measurement and no dichotomization

and

\[ -.43 < r_C < .95 \] if the correlation is corrected for measurement error and
dichotomization in both variables.

Hence, it is clear that the combination of dichotomization and measurement
error can greatly reduce the amount of information conveyed by a sample (study)
correlation.

**Imperfect Construct Validity**

**in Independent and Dependent Variables**

We define the construct validity of a measure as its true score correlation with
the actual construct or trait it is supposed to measure. The case of construct validity
is similar to the case of error of measurement if the path analysis permits a simple
multiplicative attenuation (see Chapter 2 for the required conditions). If error of
measurement is treated separately, then the impact of imperfect construct validity
is to multiply the true population correlation by an attenuation factor equal to
the construct validity of the variable. If there is imperfect construct validity in
both variables, and if both proxy variables satisfy the path analysis requirements,
then the effect is a double attenuation, that is, multiplication by the product of
the two construct validities. Because the attenuation effect is systematic, it can
be algebraically reversed. To reverse the effect of multiplying the correlation by
a constant, we divide the correlation by the same constant. Thus, to restore the
correlation to what it would have been had the variables been measured with
perfect construct validity, we divide the study correlation by the product of the two
construct validities. Note that, for perfect construct validity, we divide by 1.00,
which leaves the correlation unchanged. So perfect construct validity is a special
case of imperfect construct validity.

For example, let the construct validity of the independent variable be \( a_1 \) and
let the construct validity of the dependent variable be \( a_2 \). The impact of imperfect
construct validity is to multiply the true correlation by the product \( a_1 a_2 \). That is,

\[ \rho_o = a_1 a_2 \rho \]

The correction formula divides by the same attenuation factors

\[ \frac{\rho_o}{a_1 a_2} = \frac{(a_1 a_2 \rho)}{a_1 a_2} = \rho \]

The corrected sample correlation is thus

\[ r_C = \frac{\rho_o}{a_1 a_2} = \frac{1}{a_1 a_2} \rho_o \]

which multiplies the sampling error by the same factor \((1/a_1 a_2)\).

The attenuating effect of imperfect construct validity combines with the
attenuating effect of other artifacts. Consider another example:

\[
\begin{align*}
  a_1 &= .90 = \text{the square root of the reliability of } X; \ r_{XX} = .81, \\
  a_2 &= .90 = \text{the square root of the reliability of } Y; \ r_{YY} = .81, \\
  a_3 &= .90 = \text{the construct validity of } X, \\
  a_4 &= .90 = \text{the construct validity of } Y, \\
  a_5 &= .80 = \text{the attenuation factor for splitting } X \text{ at the median}, \\
  a_6 &= .80 = \text{the attenuation factor for splitting } Y \text{ at the median}.
\end{align*}
\]

The total impact of the six study imperfections is

\[ \rho_o = (.9)(.9)(.9)(.9)(.8)(.8)\rho = .42\rho \]

Thus, even minor imperfections add up. Had the true correlation been .50, the
study population correlation would be

\[ \rho_o = .42\rho = .42(.50) = .21 \]

a reduction of over one-half. The formula for the corrected correlation is

\[ r_C = \frac{\rho_o}{.42} = 2.38 \rho_o \]

Thus, to restore the systematic value of the study correlation, we must more than
double the correlation.
**Attrition Artifacts**

As noted in Chapter 2, attrition artifacts can usually be treated as range variation on the dependent variable. If there is range variation on the dependent variable, but not on the independent variable, then the mathematical treatment is identical to the treatment of range variation on the independent variable; we merely interchange the role of variables \( X \) and \( Y \). To our knowledge, there are no meta-analysis methods that correct for range restriction on both the independent and the dependent variables. However, as noted in Chapter 2, the approximation methods for doing this developed by Alexander, Carson, Alliger, and Carr (1987) are fairly accurate and could be adapted for use in meta-analysis. To date, this has not been done. The combined accuracy enhancements produced by use of indirect range restriction corrections (Hunter et al., 2002) and the Alexander et al. (1987) methods could be important, especially in the area of validity generalization. The more accurate validity estimates are likely to be substantially larger (Hunter et al., 2002).

**Extraneous Factors**

As discussed in Chapter 2, easily correctable extraneous factors in the research setting usually enter into the causal model for the research design in much the same form as the extraneous factors in the measurement of a proxy variable for the dependent variable. At the level of the study, it would be possible to correct the study effect correlation by partialing out the extraneous factor. If the extraneous factor was not controlled in the original analysis, then the attenuation factor is

\[
a = \sqrt{1 - \rho_{EY}^2}
\]  

(3.25)

where \( \rho_{EY} \) is the correlation between the extraneous factor and the dependent variable. The impact on the study effect correlation is to attenuate the population correlation by the multiplicative factor

\[\rho_o = a \rho\]

This formula can be algebraically reversed

\[\rho = \rho_o/a\]

to yield the correction formula for the sample correlation

\[r = r_o/a\]

The sampling error is divided by the same factor and increases correspondingly. The mathematics is otherwise identical to that for attenuation due to error of measurement. The impact of extraneous factors can combine with the impact of other artifacts. The compound impact is to multiply the attenuation factors for the other factors by the attenuation factor for the extraneous factor.
Bias in the Correlation

As noted in Chapter 2, the purely statistical bias in the sample correlation as an estimate of the population correlation is normally trivial in magnitude, and it is rarely worthwhile to correct for it. This is why this bias is not listed in Table 3.1 (or in Table 2.1 in Chapter 2). However, we provide the computations to check the size of the bias in any given application. The impact of bias is systematic and can be captured to a close approximation by an attenuation multiplier. If the population correlations are less than .70 (usually the case), then the best attenuation multiplier for meta-analysis is the linear attenuation factor (Hunter et al., 1996):

\[ a = 1 - \frac{1}{2N - 1} = \frac{2N - 2}{2N - 1} \]  
\[ (3.26) \]

This attenuation factor is most useful in meta-analysis because it is independent of the population correlation \( \rho \). For applications with population correlations larger than .70 (a rare condition), the more accurate attenuation factor is

\[ a = 1 - \frac{(1 - \rho^2)}{2N - 1} \]  
\[ (3.27) \]

Note that, if the preceding nonlinear attenuator is used, then correction for bias due to an extraneous variable should always be the last artifact corrected. In the case of multiple artifacts, the “\( \rho \)” in the attenuation formula will be the population correlation already attenuated for all other artifacts.

The impact of bias on the population correlation is a systematic reduction,

\[ \rho_o = a \rho \]

The correction formula follows from the algebraic reversal of this equation,

\[ \rho = \rho_o / a \]

to yield

\[ r_C = r_o / a \]

The sampling error is divided by the same factor and increases correspondingly. The sampling error variance is divided by \( a^2 \).

Multiple Simultaneous Artifacts

Table 3.1 lists 11 artifacts, 9 of which are potentially correctable by the use of multiplicative artifact attenuation factors. The other two artifacts are sampling error, which is corrected by a different strategy, and bad data, which can be corrected only if the bad data can be identified and corrected or thrown out (see Chapter 5). This section considers the compound of all the multiplicative artifacts. It is implicitly understood that range variation on both the independent and dependent variable is not simultaneously considered (Hunter & Schmidt, 1987b). Thus, the analysis would contain at most eight multiplicative artifacts. Note again that the artifact attenuation is caused by the real imperfections of the study design. The
Attenuation of the true correlation will thus occur whether we can correct for it or not.

Consider again the six artifact examples from the section on construct validity. The artifact attenuation factors are

\[ a_1 = .90 = \text{the square root of the reliability of } X; r_{XX} = .81, \]
\[ a_2 = .90 = \text{the square root of the reliability of } Y; r_{YY} = .81, \]
\[ a_3 = .90 = \text{the construct validity of } X, \]
\[ a_4 = .90 = \text{the construct validity of } Y, \]
\[ a_5 = .80 = \text{the attenuation factor for splitting } X \text{ at the median}, \]
\[ a_6 = .80 = \text{the attenuation factor for splitting } Y \text{ at the median}. \]

The total impact of the six study imperfections is determined by the total attenuation factor

\[ A = (.9)(.9)(.9)(.9)(.8)(.8) = .42 \]

If the true correlation for the study is \( \rho = .50 \), the attenuated study correlation is only

\[ \rho_0 = .42 \rho = .42(.50) = .21 \]

If the sample size is \( N = 26 \), then the linear bias attenuation factor is

\[ a_7 = 1 - 1/(2N - 1) = 1 - 1/25 = .96 \]

The total attenuation factor for all seven artifacts is thus

\[ A = .42(.96) = .40 \]

and the attenuated study population correlation is

\[ \rho_0 = .40 \rho = .40(.50) = .20 \]

With a sample size of 26, the sampling error variance of the uncorrected correlation is

\[ \text{Var}(\epsilon_o) = [1 - .20^2]/(26 - 1) = .0368 \]

The sampling error variance of the corrected correlation is

\[ \text{Var}(\epsilon_C) = \text{Var}(\epsilon_o)/A^2 = .0368/.40^2 = .2304 \]

Thus, the sampling error standard deviation is \( \sqrt{.2304} = .48 \). The 95% probability interval for the observed corrected correlation \( (r_C) \) is

\[ -.44 < r_C < 1.44 \]

If the sample size were increased to \( N = 101 \), the probability interval for the corrected correlation would shrink to

\[ .03 < r_C < .97 \]
The preceding example is an extreme case, but not unrealistic. All artifact values were derived from the empirical literature. This example shows that there is only limited information in the best of small-sample studies. When that information is watered down with large methodological artifacts, then there may be almost no information in the study. Ideally, studies with approximately perfect methodology would eliminate these artifacts substantively and, hence, eliminate the need for statistical correction. However, most methodological artifact values are determined by the feasibility limitations of field research. Thus, researchers frequently have little leeway for improvement. This means that there will nearly always be a need for statistical correction and, hence, there will nearly always be a need to greatly reduce sampling error.

Meta-Analysis of Individually Corrected Correlations

The examples of the preceding sections show that individual studies usually contain only very limited information. The random effects of sampling error are unavoidable. Furthermore, the other artifacts in study designs are often caused by factors outside the control of the investigator. Thus, the information in most studies is diluted by statistical artifacts such as those listed in Table 3.1, and perhaps still further by artifacts yet to be delineated and quantified in future research. Solid conclusions can thus only be built on cumulative research combining the information across studies. The traditional narrative review is clearly inadequate for this complex task (see Chapter 11). Thus, there is no alternative to meta-analysis. If one pretends that the only study artifact is sampling error, then the meta-analysis techniques given earlier in this chapter are used. However, if other artifacts are acknowledged and if information about the artifacts is available, then the values estimated by meta-analysis will be much more accurate if the study artifacts are corrected.

There are three kinds of artifacts in Table 3.1. First, there are bad data: recording, computing, reporting, and transcriptional errors. If the error is so large that the resulting correlation is an outlier in the meta-analysis, then the deviant result can be detected and eliminated (see Chapter 5). Otherwise, bad data go undetected and, hence, uncorrected. Second, there is the nonsystematic and random effect of sampling error. That effect can be eliminated in meta-analysis or at least greatly reduced. Third, the table contains nine artifacts that are systematic in nature. These we call the “correctable artifacts.” For each correctable artifact, there is a quantitative factor that must be known in order to correct the study correlation for that artifact. Given the necessary artifact values for the research domain in question, meta-analysis can correct for that artifact.

There are three cases in meta-analysis: (1) Artifact values are given in each individual study for all artifacts, (2) artifact values are only sporadically given for any artifact in the various studies, and (3) artifact values are given for each study on some artifacts but are only available sporadically on other artifacts. The case of individually available artifact information is treated in this chapter. The case of sporadically available information is covered in Chapter 4, as is the last case of mixed artifact information (i.e., the case in which artifact information is available for each study for some but not all artifacts).
We now consider the case where artifact information is available on nearly all individual studies. The missing artifact values can be estimated by inserting the mean value across the studies where information is given. This is done automatically by our VG6 computer program, described later. At that point, each artifact value is available for each study. There are then three phases to the meta-analysis: (1) performing the computations for each of the individual studies, (2) combining the results across studies, and (3) computing the estimated mean and variance of true effect size correlations in the designated research domain.

**Individual Study Computations**

The computations for each study are the computations used in correcting the correlation for artifacts. We begin with the observed study correlation and the sample size for that study. We next collect each piece of artifact information for that study. These values are placed in a table and may be read into a computer file. This analysis can be performed by the Windows-based program VG6 (described in the Appendix).

We next do the computations to correct for artifacts. Under most conditions, the effect of each correctable artifact is to reduce the correlation by an amount that can be quantified as a multiplicative factor less than 1.00, which we call the “attenuation factor.” Under these conditions, the net impact of all the correctable artifacts can be computed by simply multiplying the separate artifact attenuation factors. This results in a compound artifact attenuation factor. Dividing the observed study correlation by the compound attenuation factor corrects the study correlation for the systematic reduction caused by those artifacts.

For each study, we first compute the artifact attenuation factor for each artifact. Denote the separate attenuation factors by $a_1, a_2, a_3,$ and so forth. The compound attenuation factor for the several artifacts is the product

$$A = a_1 a_2 a_3 \ldots \quad (3.28)$$

We can now compute the corrected study correlation $r_C$. We denote the study observed correlation by $r_o$ and the corrected correlation by $r_C$. The corrected correlation is then

$$r_C = r_o / A \quad (3.29)$$

This estimate of $r_C$ has a slight downward (i.e., negative) bias (Bobko, 1983), but the degree of underestimation is very small.

For purposes of estimating sampling error, it is necessary to estimate the mean uncorrected correlation. This is the sample size weight average of the observed correlations. The sampling error variance in the corrected correlation is computed in two steps. First, the sampling error variance in the uncorrected correlation is computed. Then the sampling error variance in the corrected correlation is computed from that. The sampling error variance in the uncorrected correlation, $\text{Var}(e_o)$, is

$$\text{Var}(e_o) = [1 - \bar{r}_o^2]^2 / (N_i - 1)$$
where $\bar{r}_o$ is the mean uncorrected correlation across studies (Hunter & Schmidt, 1994b; Law et al., 1994b) and $N_i$ is the sample size for the study in question. The sampling error variance in the corrected correlation is then given by

$$\text{Var}(e_C) = \text{Var}(e_o)/A^2$$

where $A$ is the compound artifact attenuation factor for that study. For simplicity, denote the sampling error variance by $ve$. That is, define $ve$ by

$$ve = \text{Var}(e_C)$$

The preceding computation of sampling error variance is corrected for all artifacts. However, we can refine our estimate of the contribution of the range correction to sampling error. Because the attenuation factor for range variation contains the correlation itself, the corresponding sampling error increase is only in proportion to the derivative of the attenuation instead of the attenuation factor itself. This difference is small in many cases. However, a more accurate estimate can be computed using a more complicated formula for sampling error variance. Compute a first estimate of $ve$ using the preceding formula. Label the first estimate $ve'$. The improved estimate is then

$$ve = a^2 ve'$$

where $a$ is computed as

$$a = 1/[(U_X^2 - 1)r_o^2 + 1]$$

in the case of direct range restriction. In the case of indirect range restriction, use the same formula for $a$ but substitute $U_T$ for $U_X$. Consider an extreme example: Suppose only the top half of the ability distribution had been selected in a personnel selection study (i.e., direct range restriction). The study population standard deviation is smaller than the reference population standard deviation by the factor $u_X = .60$. The reciprocal of $u_X$ would be $U_X = 1/.60 = 1.67$. If the study correlation ($r_o$) were .20, then $a = .94$, and the refining factor would be $a^2 = .88$. That is, the sampling error variance for that study would be 12% smaller than estimated using the simple attenuation factor. (See Bobko, 1983, for an equation for $ve$ that yields values essentially identical to the equation for $ve$ presented here.) Thus, for each study $i$, we generate four numbers: the corrected correlation $r_{ci}$, the sample size $N_i$, the compound attenuation factor $A_i$, and the sampling error variance $ve_i$. These are the numbers used in the meta-analysis proper.

**Combining Across Studies**

Meta-analysis reduces sampling error by averaging sampling errors. Thus, an important step in meta-analysis is the computation of certain critical averages: the average correlation, the variance of correlations (the average squared deviation from the mean), and the average sampling error variance. To do this, we must decide how much weight to give to each study.
Meta-Analysis of Correlations Corrected Individually for Artifacts

What weights should be used? The first step in the meta-analysis is to average certain numbers across studies. This averaging can be done in several ways. In averaging corrected correlations, a simple or “unweighted” average gives as much weight to a study with a sample size of 12 as to a study with a sample size of 1,200. Yet the sampling error variance in the small-sample study is 100 times greater than that in the large-sample study. Schmidt and Hunter (1977) noted this problem and recommended that each study be weighted by its sample size. Hunter, Schmidt, and Jackson (1982, pp. 41–42) noted that this is an optimal strategy when there is little or no variation in population correlations across studies—the “homogeneous” case (Hunter & Schmidt, 2000). They noted that there can be a problem if correlations differ a great deal across studies. The problem is potentially acute if the meta-analysis contains one study with very large sample size while all the other studies have much smaller sample size. If the large-sample-size study is deviant in some way, then, because it dominates the meta-analysis, the meta-analysis will be deviant. This case arises rarely in practice.

The homogeneous case was considered in technical detail in Hedges and Olkin (1985, Chapter 6). They noted the key mathematical theorem for that case. If the population correlations do not differ from one study to the next, then the optimal weights are obtained by weighting each study inversely to its sampling error variance. In the case of correlations that are not corrected for any artifacts, the sampling error variance is

$$\text{Var}(e_i) = \frac{(1 - \rho_i^2)^2}{(N_i - 1)}$$

The optimal weight for this homogeneous case would thus be

$$w_i = \frac{(N_i - 1)}{(1 - \rho_i^2)^2}$$

Because the population correlation $\rho_i$ is not known, this optimal weight cannot be used. Although Hedges and Olkin (1985) did not note it, even in the homogeneous case, the substitution of the observed study correlation $r_{oi}$ for the study population correlation $\rho_i$ does not lead to the most accurate alternative weights. As noted earlier in this chapter, a more accurate alternative is to substitute the mean observed correlation $\bar{r}_o$ for each study population $\rho_i$ (Hunter & Schmidt, 1994b; Law et al., 1994b; see Chapter 5). Because the resulting multiplicative term

$$1/(1 - \bar{r}_o^2)^2$$

is the same for all studies, it can be dropped. That is, the corresponding weighting can be more easily accomplished by using the weight

$$w_i = N_i - 1$$

As Hedges and Olkin noted, this differs only trivially from the Schmidt and Hunter (1977) recommendation to weight each study by sample size:

$$w_i = N_i$$

(3.31)

The previous discussion is cast in terms of the homogeneous case (i.e., the case where $S_0^2 = 0$). However, as noted earlier in this chapter, these weights are also
very accurate for the heterogeneous case (where $S^2_\rho > 0$), so long as the sample size is not correlated with the size of $\rho_i$ (the population correlation). Except for that rare situation, weighting by $N_i$ is very accurate for the heterogeneous case as well as for the homogeneous case (see, e.g., Hall & Brannick, 2002). This is important because some have maintained that one should not weight by $N_i$ in the heterogeneous case. (See the discussion of fixed- vs. random-effects meta-analysis models in Chapters 5 and 9.)

When studies differ greatly on one or more of the artifacts corrected, then more complicated weighting will make better use of the information in the studies. Studies with more information should receive more weight than studies with less information. For example, studies in which one or both of the variables is dichotomized with extreme splits should receive much less weight than a study with a near-even split. The same is true if one study has very low reliability while a second study has high reliability.

The sampling error variance of the corrected correlation is

$$\text{Var}(e_C) = \frac{\left(1 - \rho_i^2\right)^2}{(N_i - 1)/A_i^2}$$

Thus, the weight for each study would be

$$w_i = A_i^2 \left(\frac{(N_i - 1)\rho_i^2}{1 - \rho_i^2}\right)$$

$$= \left(\frac{(N_i - 1)A_i^2}{1 - \rho_i^2}\right)^2$$

Because the study population correlation $\rho_i$ is not known, some substitution must be made. As noted previously, the most accurate substitution is to substitute the mean observed correlation $\bar{r}_o$ for each study population correlation $\rho_i$ (Hunter & Schmidt, 1994b; Law et al., 1994b). Because $\bar{r}_o$ is a constant, this has the effect of eliminating the term with $\rho_i$ and thus yields the weights

$$w_i = (N_i - 1)A_i^2$$

This, in turn, differs only trivially from the simpler weights

$$w_i = N_iA_i^2$$  \hspace{1cm} (3.32)

That is, the weight for each study is the product of two factors: the sample size $N_i$ and the square of the artifact attenuation factor $A_i$. The attenuation factor is squared because to multiply a correlation by the factor $1/A_i$ is to multiply its sampling error variance by $1/A_i^2$. This weighting scheme has the desired effect: The more extreme the artifact attenuation in a given study, the less the weight assigned to that study. That is, the more information contained in the study, the greater its weight.

Consider two studies with sample size 100. Assume that (1) both variables are measured with perfect reliability and without range restriction, (2) the population correlation $\rho$ is the same in both studies, and (3) the dependent variable is dichotomized in both studies. In Study 1, there is a 50–50 split so the true population correlation $\rho$ is reduced to a study population correlation of $.80 \rho$. In Study 2, there is a 90–10 split so the true population correlation $\rho$ is reduced to a study population correlation of $.59 \rho$. To correct for the attenuation due to artificial
dichotomization, the Study 1 observed correlation $r_{o1}$ must be multiplied by the reciprocal of .80, that is, $1/.80 = 1.25$:

$$r_1 = 1.25r_{o1}$$

Thus, the sampling error variance is multiplied by the square of 1.25, that is, $1.25^2 = 1.5625$. To correct for the attenuation due to artificial dichotomization, the Study 2 observed correlation $r_{o2}$ must be multiplied by the reciprocal of .59, that is, $1/.59 = 1.695$:

$$r_2 = 1.695r_{o2}$$

Thus, the sampling error variance is multiplied by the square of 1.695, that is, $1.695^2 = 2.8730$. Before the dichotomization and the correction, the sampling error in both correlations was the same: the sampling error implied by equal sample sizes of 100 and equal population correlations of $\rho_i$. However, after the correction for dichotomization, the second study has sampling error variance that is $2.8730/1.5625 = 1.839$ times larger than the sampling error variance in the first study. Thus, the second study deserves only $1/1.839 = .544$ times as much weight in the meta-analysis. The weights assigned to the studies by the formula

$$w_i = N_iA_i^2$$
do just that:

$$w_1 = 100(.80)^2 = 100(.64) = 64$$
and

$$w_2 = 100(.59)^2 = 100(.35) = 35$$

where $64/35 = 1.83$. Thus, to use round numbers, the study with twice the information is given twice the weight (1.00 vs. .544).

In summary, there are three typical types of weights that might be used in averaging. First, one could ignore the differences in quality (information content) between studies and give as much weight to high-error studies as to low-error studies. This is the case of equal weights and is represented by $w_i = 1$. Second, one might consider the effect on quality of unequal sample size but ignore the effects of other artifacts. This leads to the sample size weights recommended by Schmidt and Hunter (1977): $w_i = N_i$. Third, one might consider the weights that take into account both $N_i$ and other artifacts: $w_i = N_iA_i^2$.

We recommend the use of these last weights, because they give less weight to those studies that require greater correction and, hence, have greater sampling error. These weights are used in our Windows-based program for meta-analysis of correlations corrected individually (described in the Appendix).

**Final Meta-Analysis Estimation**

Once the researcher has corrected each study correlation for artifacts and has decided on the weights, there are three meta-analysis averages to be computed using the corrected $r$s:

$$\bar{r}_C = \frac{\sum w_i r_{Ci}}{\sum w_i}$$

(3.33)
Var(r_C) = \Sigma w_i [r_{C_i} - \bar{r}_C]^2 / \Sigma w_i \hspace{1cm} (3.34)

\text{Ave}(ve) = \Sigma w_i ve_i / \Sigma w_i \hspace{1cm} (3.35)

The mean actual correlation is then estimated by the mean corrected correlation:

\bar{\rho} = \bar{r}_C \hspace{1cm} (3.36)

The variance of actual correlations is given by the corrected variance of corrected correlations:

\text{Var}(\rho) = \text{Var}(r_C) - \text{Ave}(ve) \hspace{1cm} (3.37)

The standard deviation is estimated by the square root of the variance estimate if it is positive. If the standard deviation is actually 0, then the variance estimate will be negative half the time by chance. As discussed earlier in this chapter, a negative variance estimate suggests that the actual variance is 0 (see also Chapter 9).

\textbf{An Example: Validity Generalization}

We will now consider a hypothetical, but realistic validity generalization example for a test of verbal aptitude. The reliability of this test computed in the applicant group (unrestricted group) is $r_{XX_a} = .85$. The 12 studies shown in Table 3.5a have been conducted on job incumbents who have earlier been hired using a variety of procedures of which we have no record; hence, any range restriction would be indirect. The observed validities range from .07 to .49, and only 7 of the 12 validity estimates are statistically significant. The $u_X$ values in the second column in Table 3.5a are all less than 1.00, indicating there has been indirect range restriction. Using Equation (3.16), we convert these $u_X$ values into $u_T$ values, which are shown in the table. The first column of job performance reliabilities ($r_{yy}$) represents the restricted values (i.e., those computed in the incumbent samples). These are the $r_{yy}$ values that are needed for the computations in this example. The $r_{yy}$ values—the estimated criterion reliabilities in the applicant group—are presented for only general informational purposes. (These values are computed using Equation [5.20] in Chapter 5.) The test reliabilities in Table 3.5a are the restricted (incumbent group) values (i.e., $r_{XX_i}$). Note that, because of the effects of the indirect range restriction, the $r_{XX_i}$ values are less than the applicant pool value of $r_{XX_a} = .85$. (Test reliabilities are computed from the $r_{XX_i} = .85$ value using Equation [3.17c].) We first correct each observed correlation for measurement error in both variables and then correct for range restriction, using the $u_T$ values in the range restriction correction formula. These corrected correlations are shown in the last column in Table 3.5a. The underlying true score (construct level) correlation used to generate these hypothetical data is $\rho = .57$. However, the true score correlation overestimates operational validity, because in actual test use we must use observed test scores to predict future job performance and cannot use applicants’ (unknown) true scores. Hence, the underlying true validity used to generate this example is $\rho_{xy_t} = \sqrt{.85(.57)} = .53$. In the case of indirect range restriction, meta-analysis must correct for measurement error in both variables.
A Worked Example: Indirect Range Restriction

We now proceed with the meta-analysis of the data in our example. Our calculations are the same as those in the computer program VG6-I (for indirect range restriction). (See the Appendix for availability of the software package containing this program.) We weight each study as follows:

\[ w_i = N_i A_i^2 \]

The worksheet for these calculations is shown in Table 3.5b. The attenuation factor for each study is most easily computed as the ratio of the uncorrected correlation to the corrected correlation. For Study 1,

\[ A_1 = \frac{.35}{.80} = .43 \]

Thus, the study weight shown in the next to the last column is

\[ w_1 = 68(.43)^2 = 68(.18) = 12.6 \]

The sample size weighted-average uncorrected sample correlation is \( \bar{r} = .28 \). Thus, the sampling error variance of the uncorrected correlation in Study 1 is

\[ \text{Var}(e) = (1 - .28^2)/67 = .012677 \]

Because all studies in this hypothetical example have the same sample size, this is the estimate of sampling error variance in the uncorrected correlations for every study (not shown in Table 3.5b).

The sampling error variance in the corrected correlations can be estimated in two ways. We could use the simple estimate that ignores the nonlinearity of the range restriction correction:

\[ \text{Var}(ve) = \text{Var}(e)/A^2 \]

For Study 1, this value is

\[ \text{Var}(ve) = .012677/(.43)^2 = .068561 \]

This estimate is recorded in the column “Simple Error Variance” in Table 3.5b. However, a more accurate estimate is obtained by using the correction factor computed using the derivative of the correction transformation:

\[ a = 1/[(U_T^2 - 1)r_o^2 + 1] \]
For Study 1, the standard deviation ratio is

\[ U_T = 1/u_T = 1/1.468 = 2.1368 \]

So the correction factor is

\[ a = 1/[(2.1368^2 - 1)(.35)^2 + 1] = 1/1.4368 = .6960 \]

Thus, the refined estimate of sampling error variance in Study 1 is

\[ \text{Var}(ve) = a^2(.068561) = .033212 \]

as shown in the column “Refined Error Variance” in Table 3.5b.

Again, we note that the corrected correlations are denoted by \( r_C \). The three weighted averages are

\[ \hat{\rho} = \bar{r}_C = .574 \]

\[ \text{Var}(r_C) = .034091 \]

\[ \text{Var}(e) = .048120 \text{ (Simple method)} \]

\[ \text{Var}(e) = .035085 \text{ (Refined method)} \]

The estimated mean true score correlation (.574) is very close to the actual value (.570). Because all population correlations were assumed to be the same, the variance of population correlations is actually 0. Thus, the sampling error variance should equal the observed variance. For the simple estimation method, we have

\[ \text{Var}(\rho) = \text{Var}(r_C) - \text{Var}(e) = .034091 - .048120 = -.014029 \]

which correctly suggests a standard deviation of 0 but which is a poorer estimate than we would like. For the refined estimation method, we have

\[ \text{Var}(\rho) = \text{Var}(r) - \text{Var}(e) = .034091 - .035085 = -.000994 \]

The value \(-.000994\) is much closer to the true value of 0 and, again, correctly indicates that the true value is 0.

Meta-analysis correcting each correlation individually can be carried out on the computer using the program VG6. (See the Appendix for a description of the software package.) This program has separate subroutines for direct and indirect range restriction. In the case of indirect range restriction, as in our example here, the reliabilities entered into the program should be those in the restricted group. (If unrestricted independent variable reliabilities are entered, the program will convert them to restricted group values, using Equation [3.17].) Reliability corrections for both variables are made prior to the correction for range restriction (see Chapter 5 and Hunter et al., 2002). If range restriction is direct, the reliability correction for the dependent variable can be made prior to the range restriction correction, but the correction for unreliability in the independent variable must be made after the range restriction correction, using the reliability \( r_{XX} \) in the unrestricted population. The reason for this is that direct range restriction on the independent variable causes true scores and measurement errors to be (negatively) correlated in the selected sample, hence violating the critical assumption underlying reliability
Table 3.5a Hypothetical validity and artifact information

<table>
<thead>
<tr>
<th>Study</th>
<th>sx/Sx (µX)</th>
<th>sy/Sy (µT)</th>
<th>Reliability (rXX)</th>
<th>Criterion Reliability (rY)</th>
<th>Sample Size</th>
<th>Observed Correlation</th>
<th>Corrected Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.580 .468</td>
<td>.55</td>
<td>80 .86</td>
<td>68 .35**</td>
<td></td>
<td>.80</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.580 .468</td>
<td>.55</td>
<td>60 .61</td>
<td>68 .07</td>
<td></td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.580 .468</td>
<td>.55</td>
<td>80 .81</td>
<td>68 .11</td>
<td></td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.580 .468</td>
<td>.55</td>
<td>60 .70</td>
<td>68 .31*</td>
<td></td>
<td>.81</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.678 .603</td>
<td>.67</td>
<td>80 .81</td>
<td>68 .18</td>
<td></td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.678 .603</td>
<td>.67</td>
<td>60 .67</td>
<td>68 .36**</td>
<td></td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.678 .603</td>
<td>.67</td>
<td>80 .84</td>
<td>68 .40**</td>
<td></td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.678 .603</td>
<td>.67</td>
<td>60 .61</td>
<td>68 .13</td>
<td></td>
<td>.33</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.869 .844</td>
<td>.80</td>
<td>80 .82</td>
<td>68 .49**</td>
<td></td>
<td>.68</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.869 .844</td>
<td>.80</td>
<td>60 .61</td>
<td>68 .23</td>
<td></td>
<td>.38</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>.869 .844</td>
<td>.80</td>
<td>80 .81</td>
<td>68 .29*</td>
<td></td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>.869 .844</td>
<td>.80</td>
<td>60 .63</td>
<td>68 .44*</td>
<td></td>
<td>.70</td>
<td></td>
</tr>
</tbody>
</table>

** p < .01 (two-tailed).
* p < .05 (two-tailed).

NOTE: rY/Y, the criterion reliability in the applicant group, is not used in the example. It is shown only for informational purposes.

Table 3.5b Meta-analysis worksheet

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Size</th>
<th>Attenuation Factor</th>
<th>Simple Error Variance</th>
<th>Refined Error Variance</th>
<th>Study Weight</th>
<th>Corrected Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>.43</td>
<td>.0686</td>
<td>.0332</td>
<td>12.6</td>
<td>.80</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>.27</td>
<td>.1739</td>
<td>.0863</td>
<td>5.0</td>
<td>.26</td>
</tr>
<tr>
<td>3</td>
<td>68</td>
<td>.32</td>
<td>.1238</td>
<td>.1138</td>
<td>7.0</td>
<td>.34</td>
</tr>
<tr>
<td>4</td>
<td>68</td>
<td>.38</td>
<td>.0878</td>
<td>.0488</td>
<td>9.8</td>
<td>.81</td>
</tr>
<tr>
<td>5</td>
<td>68</td>
<td>.47</td>
<td>.0574</td>
<td>.0319</td>
<td>15.0</td>
<td>.39</td>
</tr>
<tr>
<td>6</td>
<td>68</td>
<td>.48</td>
<td>.0550</td>
<td>.0365</td>
<td>15.7</td>
<td>.75</td>
</tr>
<tr>
<td>7</td>
<td>68</td>
<td>.55</td>
<td>.0419</td>
<td>.0256</td>
<td>20.6</td>
<td>.73</td>
</tr>
<tr>
<td>8</td>
<td>68</td>
<td>.39</td>
<td>.0833</td>
<td>.0785</td>
<td>10.3</td>
<td>.33</td>
</tr>
<tr>
<td>9</td>
<td>68</td>
<td>.72</td>
<td>.0244</td>
<td>.0203</td>
<td>35.3</td>
<td>.68</td>
</tr>
<tr>
<td>10</td>
<td>68</td>
<td>.60</td>
<td>.0352</td>
<td>.0337</td>
<td>24.5</td>
<td>.38</td>
</tr>
<tr>
<td>11</td>
<td>68</td>
<td>.69</td>
<td>.0266</td>
<td>.0233</td>
<td>32.3</td>
<td>.42</td>
</tr>
<tr>
<td>12</td>
<td>68</td>
<td>.63</td>
<td>.0319</td>
<td>.0274</td>
<td>27.0</td>
<td>.70</td>
</tr>
</tbody>
</table>

NOTE: Figures in this table apply to the corrected correlations.

Meta-Analysis of Correlations Corrected Individually for Artifacts

Table 3.5 A meta-analysis on hypothetical personnel selection studies with indirect range restriction

Theoretical considerations (see Chapter 5; Hunter et al., 2002; Mendoza & Mumford, 1987). Under these conditions, the reliability of the independent variable in the selected sample is undefined and undefinable; hence, this correction must be made after the range restriction correction has been made, using an estimate of $r_{XX_2}$ obtained in the unrestricted sample. Mendoza and Mumford (1987) were the first to point out this important fact. Hunter et al. (2002) incorporated this consideration into their meta-analysis procedures. Of course, if the meta-analysis is a validity generalization study, there is no need to correct for unreliability in the independent variable,
and this correction is simply omitted. (An example of a validity generalization meta-analysis correcting correlations individually under conditions of direct range restriction is presented in Chapter 3 of the 1990 edition of this book.) For further details, see Chapter 5.

Returning to our example, the mean corrected correlation of .57 that we obtained does not estimate the true validity, because it has been corrected for measurement error in the independent variable (the predictor test). The value .57 is the mean true score correlation, not the mean operational validity \( \rho_{xy} \). Because we know the reliability of this test in the applicant pool (unrestricted) group (i.e., \( r_{XX} = .85 \), as noted at the beginning of our example), we can easily calculate the needed operational validity estimates as follows:

\[
\rho_{xy} = \sqrt{r_{XX}} \rho
\]

\[
\rho_{xy} = \sqrt{.85}(.574) = .53
\]

\[
SD_{\rho_{xy}} = \sqrt{.85} SD_{\rho}
\]

\[
SD_{\rho_{xy}} = \sqrt{.85}(0) = 0
\]

Again, both these values are correct. Although we need the operational validity estimate for applied work in personnel selection, it is the construct-level correlation of .57 that is needed for theoretical research (theory testing). Most meta-analyses outside the area of personnel selection are theoretical in orientation.

Next, for comparison purposes, we present sampling error corrections for the observed correlations (bare-bones meta-analysis) as well as for the fully corrected correlations (complete meta-analysis). The comparative results are as follows:

<table>
<thead>
<tr>
<th>Uncorrected Correlations</th>
<th>Fully Corrected Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{r}_{o} = .28 )</td>
<td>( \bar{r} = .57 )</td>
</tr>
<tr>
<td>( \sigma^2_{r_o} = .017033 )</td>
<td>( \sigma^2 = .034091 )</td>
</tr>
<tr>
<td>( \sigma^2_{\rho} = .012677 )</td>
<td>( \sigma^2 = .035085 )</td>
</tr>
<tr>
<td>( \sigma^2_{\rho} = .004356 )</td>
<td>( \sigma^2 = -.000994 )</td>
</tr>
<tr>
<td>( SD_{\rho} = .07 )</td>
<td>( SD_{\rho} = .0 )</td>
</tr>
</tbody>
</table>

These results may be checked using the Windows-based program VG6-I (see the Appendix for availability). Program results are the same as those calculated here. The results for uncorrected correlations show an incorrect mean of .28 (vs. the true value of .57) and a standard deviation of .07 (vs. an actual standard deviation of 0). The bare-bones meta-analysis is very inaccurate in comparison to the actual effect size correlation used to generate these hypothetical data. However, the error in the uncorrected correlation meta-analysis is exactly consistent with the known artifacts. The average attenuation factor is .494 (using the weights that would be used for the meta-analysis of uncorrected correlations, i.e., weighting by sample size). The mean correlation would then be expected to be reduced from .57 to .494(.57) = .282, which matches the mean uncorrected correlation. A check of Table 3.5b shows considerable variation in the attenuation factors due to variation in the artifact values across studies. In the bare-bones meta-analysis of the standard deviation, this variation of .07 across studies is all artifactual; that is, even after
correction for sampling error, the uncorrected correlations vary across studies due to differences in range restriction and criterion unreliability.

If the variance beyond sampling error in uncorrected correlations is interpreted to mean that there are real moderator variables present, that is a major substantive error of interpretation. However, if it is correctly noted that no correction for artifacts other than sampling error was made and that the remaining variation might be due to such artifacts, then the substantive error would be avoided. The moral of this story is this: Even meta-analysis will not save a review from critical errors if there is no correction for study artifacts beyond sampling error. Such artifacts are present in virtually every research domain.

This method of meta-analysis, in which each study correlation is individually corrected for attenuating artifacts, can often not be used. The reason is obvious: Many study sets fail to provide all the needed artifact information. For examples of large-scale meta-analyses in which this method could be and was applied, see Carlson et al. (1999) and Rothstein et al. (1990). These studies were both consortium validation efforts in which numerous employers participated. Quite a number of such consortium studies have been conducted over the last 15 to 20 years (e.g., Dunnette et al., 1982; Dye, 1982; Peterson, 1982). The meta-analysis methods described in this chapter could be applied to the data from these studies. Another example of a meta-analysis in which correlations were corrected individually is the study by Judge et al. (2001) relating job satisfaction to job performance.

When there is no range restriction (and, hence, no corrections for range restriction), meta-analysis correcting each correlation individually is much simpler than the example presented here. When there is no range restriction, sampling error and measurement error are often the only artifacts for which corrections can be made. In such cases, one need not consider the impact of range restriction on the reliabilities. That is, the distinction between reliabilities in the restricted and unrestricted groups does not exist. Also, there is no need for the special step used in our example to compute the “refined” estimate of sampling error variance for the corrected correlations. Hence, such meta-analyses are much easier to compute using a calculator. However, for these simpler meta-analyses, the same computer program (VG6) can still be used. In terms of data input to the program, the only difference is that all values of $u_X$ or $u_T$ are entered as 1.00 (indicating no range restriction). The program output presents the same final estimates (e.g., $\bar{\rho}$ and $SD_\rho$) as it does for the case in which there is range restriction.

In addition to the methods described in this chapter, Raju, Burke, Normand, and Langlois (1991) also developed procedures for conducting meta-analysis on correlations corrected individually for artifacts. Their procedures take into account sampling error in the reliability estimates. These procedures have been incorporated into a computer program called MAIN (Raju & Fleer, 1997), which is available from Nambury Raju. At present, this program is set up to assume that range restriction is direct, but it could be modified in the future to incorporate the corrections for indirect range restriction described in this chapter. As of this writing, the computer program Comprehensive Meta-Analysis (published by Biostat; www.metaAnalysis.com) does not contain a subprogram for meta-analysis of individually corrected correlations. However, it is planned that this feature will
be added in later versions of the program. Further discussion of available software for meta-analysis can be found in Chapter 11.

Summary of Meta-Analysis
Correcting Each Correlation Individually

Correlations are subject to many artifactual sources of variation that we can control in a meta-analysis: the random effect of sampling error and the systematic attenuation produced by correctable artifacts such as error of measurement, dichotomization, imperfect construct validity, or range variation. Sampling error and error of measurement are found in every study; hence, a full meta-analysis should always correct for both. Some domains are not subject to significant amounts of range variation across studies, but in areas with high degrees of subject selection, such as personnel selection research, the effects of range restriction can be as large as the effects of error of measurement. The attenuation produced by dichotomization is even larger than the effects produced by error of measurement in most domains. (However, measurement error is always present while dichotomization is often not present.) Failure to correct for these artifacts results in massive underestimation of the mean correlation. Failure to control for variation in these artifacts results in a large overstatement of the variance of population correlations and, thus, in a potential false assertion of moderator variables where there are none.

All meta-analyses can be corrected for sampling error. To do this, we need only know the sample size $N_i$ for each sample correlation $r_i$. Meta-analysis correcting for only sampling error has come to be called “bare-bones” meta-analysis. However, in this analysis, the correlations analyzed are the correlations between imperfectly measured variables (which are therefore systematically lowered by error of measurement); the correlations may be computed on observed rather than reference populations (so that these correlations are biased by range variation); the correlations may be greatly attenuated by dichotomization; and the correlations may be much smaller than they would have been had the construct validity in each study been perfect. Thus, the mean correlation of a bare-bones meta-analysis is a biased estimate of the desired mean correlation, that is, the correlation from a study conducted without the imperfections that stem from limited scientific resources (Rubin, 1990). Furthermore, although the variance in a bare-bones meta-analysis is corrected for sampling error, it is still biased upward because it contains variance due to differences in reliability, differences in range (if any), differences in extremity of split in dichotomization (if any), and differences in construct validity. Thus, the “bare-bones” variance is usually a very poor estimate of the real variance.

For any artifact, if the needed artifact information is available from each study, then each correlation can be separately corrected for that artifact. The corrected correlations can then be analyzed by meta-analysis to eliminate sampling error. This chapter presented detailed procedures for this form of meta-analysis: meta-analysis of correlations corrected individually for the effects of artifacts.
In most sets of studies to be subjected to meta-analysis, such complete artifact information will not be available. Such study sets do occur, however, and may become more frequent in the future as reporting practices improve. In three cases to date known to us where this information has been available (Carlson et al., 1999; Dye, 1982; Rothstein et al., 1990), this form of meta-analysis has found virtually all between-study variance in correlations to be due to artifacts. However, if the fully corrected variance of correlations across studies is far enough above 0 to suggest that there is a moderator variable, then appropriate potential moderator variables can be checked by analyzing subsets of studies (as was done in Judge et al., 2001). That is, a separate meta-analysis is conducted within each subset of studies. Alternatively, study characteristics that are potential moderators can be coded and correlated with study correlations. This chapter described methods for both these approaches to moderator analysis. See also the last section in Chapter 8.

In many sets of studies, information on particular artifacts is available in some studies, but not in others. Because some (often much) artifact information is missing, it is not possible to fully correct each study correlation for the attenuating effects of every artifact. Nevertheless, it is possible to conduct accurate meta-analyses that correct the final meta-analytic results for all artifacts. This is accomplished using distributions of artifact effects compiled from studies that do provide information on that artifact. These methods of meta-analysis are the subject of the next chapter.
Exercise 3.1

Bare-Bones Meta-Analysis: Correcting for Sampling Error Only

These are real data. Psychological Services, Inc. (PSI), of Los Angeles, a selection consulting company, conducted a consortium study that 16 companies participated in. These firms were from all over the United States and from a variety of industries.

The job was mid-level clerical work, and one scale studied was the composite battery score of four clerical tests. The criterion was ratings of job performance based on rating scales specially developed for this study. The same scales were used in all 16 companies.

Following are the sample sizes and observed correlations. Conduct a meta-analysis that corrects only for sampling error (i.e., a bare-bones meta-analysis). Remember to use $\bar{r}$ in the formula for sampling error variance. Give and explain the following:

1. The mean observed correlation.
2. The observed SD and variance.
3. The expected sampling error variance.
4. The corrected SD and variance.
5. The percentage of variance accounted for by sampling error.
6. What observed SD would you expect for these correlations if all observed variance were due to sampling error?
7. What additional corrections need to be made (beyond those that are part of this exercise) before this meta-analysis could be considered complete? Why? What effects would these corrections correct for?

For maximum learning, you should carry out this exercise using a calculator. However, it can also be carried out using either of our computer programs for correlations corrected individually (VG6-D and VG6-I). Both programs provide bare-bones meta-analysis results. See the Appendix for a description of these programs.

<table>
<thead>
<tr>
<th>Company</th>
<th>$N$</th>
<th>Observed Validity of Sum of the Four Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>203</td>
<td>.16</td>
</tr>
<tr>
<td>2</td>
<td>214</td>
<td>.19</td>
</tr>
<tr>
<td>3</td>
<td>225</td>
<td>.17</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>.04</td>
</tr>
<tr>
<td>5</td>
<td>34</td>
<td>.13</td>
</tr>
<tr>
<td>6</td>
<td>94</td>
<td>.38</td>
</tr>
<tr>
<td>7</td>
<td>41</td>
<td>.36</td>
</tr>
<tr>
<td>8</td>
<td>323</td>
<td>.34</td>
</tr>
<tr>
<td>9</td>
<td>142</td>
<td>.30</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>.20</td>
</tr>
<tr>
<td>11</td>
<td>534</td>
<td>.36</td>
</tr>
<tr>
<td>12</td>
<td>30</td>
<td>.24</td>
</tr>
<tr>
<td>13</td>
<td>223</td>
<td>.36</td>
</tr>
<tr>
<td>14</td>
<td>226</td>
<td>.40</td>
</tr>
<tr>
<td>15</td>
<td>101</td>
<td>.21</td>
</tr>
<tr>
<td>16</td>
<td>46</td>
<td>.35</td>
</tr>
</tbody>
</table>
Meta-Analysis of Correlations Corrected Individually for Artifacts

Exercise 3.2

Meta-Analysis Correcting Each Correlation Individually

Find the following journal article in a library and make a personal copy of it:


Brown analyzed his data as if it were unmatched. That is, even though he had a criterion reliability value and a range restriction value specifically for each observed validity coefficient, he did not correct each observed coefficient individually. Instead, he used an unmatched data approach; that is, he used artifact distribution meta-analysis. (See Chapter 4.) In fairness to him, at the time he conducted his analysis, we had not yet published a description of methods for correcting each correlation individually.

Assume that range restriction in his data is direct and conduct new meta-analyses of his data correcting each observed validity individually. In the first meta-analysis, weight each study only by sample size (original weighting method). (This analysis cannot be run using the VG6-D program, because it does not weight by \( N_i \).) Then do the meta-analysis again, weighting each study by \( N_i A_i^2 \) as described in this chapter. The analysis weighting by \( N_i A_i^2 \) can be done using the program VG6-D (i.e., the program for correcting correlations individually; specify that range restriction is direct). This program is described in the Appendix. For each set of analyses, report the following separately for the Group A and Group B companies and for the combined data:

1. Mean true validity
2. Standard deviation of true validities
3. Value at the 10th percentile
4. Percentage variance accounted for
5. Observed standard deviation of corrected validities
6. Standard deviation predicted from artifacts
7. Number of studies (companies)
8. Total \( N \) across the studies

Present these items in columns 1 through 8, in the order given here (with the columns labeled). (Each meta-analysis is then one row of your table.)

Compare your values to those obtained by Brown in each case. (Note: Brown did not compute values at the 10th percentile, but you can compute values for his data from information he gives in the study. Be careful to use the correct \( SD_\rho \) in this calculation: Use \( SD_\rho \) for true validity, not \( SD_\zeta \) for true score correlations.) How different are the conclusions reached when these two different methods of meta-analysis are applied to the same data? What effect does weighting each study by other artifact indexes in addition to sample size have on the results? Why do you think this is the case? Do your results support Brown’s original conclusions?
## Worksheet–Exercise 3.2

Weighted by Sample Size

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Weighted by $N_i A_i^2$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data From Brown (1981)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>