This chapter presents a substantive discussion of the evaluation of experiments and interventions. The next chapter (Chapter 7) will present the quantitative methods and formulas for meta-analysis and other more technical material. For purposes of simplicity, we will consider only a two-group experiment. The principles developed here apply equally to more complicated designs.

This presentation will parallel that for correlational studies in Chapter 2. For typical studies, sampling error causes error in treatment effects and causes studies to appear to be inconsistent with each other. If the usual analysis were based on confidence intervals, the large effects of sampling error would be recognized, and spurious differences between studies would be properly attributed to sampling error. Instead, most investigators rely on the statistical significance test, which aggravates rather than reduces the problem. Meta-analysis can disentangle differences due to sampling error from differences due to real moderator variables. Treatment effects are also distorted by other artifacts: error of measurement in the dependent variable, error of measurement in the treatment variable (i.e., differences between the nominal treatment and the actual treatment), dichotomization of a continuous dependent variable, range variation on the dependent variable, lack of perfect construct validity in the dependent variable, lack of perfect construct validity in the treatment variable (e.g., confounding of the intended treatment impact with other unintended impacts), bias in the estimation of the treatment effect, as well as bad data due to reporting errors, computation errors, transcription errors, and so on.

The distortions in treatment effects produced by artifacts were camouflaged by the traditional dichotomous description of treatment effects as either “had an effect” or “had no effect.” Most artifacts reduce the size of the treatment effect. Had there been no effect to reduce, the artifact would cause no distortion. Thus, under the null hypothesis of “no effect,” artifacts other than sampling error become irrelevant and were traditionally ignored. However, meta-analysis has shown that
the nihilistic null hypothesis is rarely true. For example, as discussed in Chapters 2, 3, and 5, Lipsey and Wilson (1993) examined available meta-analyses of over 300 psychological interventions (treatment conditions) and found that only two of the treatments (less than 1%) had essentially no effect. Based on this massive study, one would estimate the prior probability that the null hypothesis is false in studies of psychological treatments at .993. In most research domains, the null hypothesis is not true and the reduction in an effect by artifacts has a real and important effect. Among other things, reduction in the size of the study effect by an artifact increases the error rate of the conventional statistical significance test (which is high in the best of conditions for most studies). Differences in the extent of artifacts between studies cause apparent differences in effects across studies, that is, produce the appearance of situation (or setting) by treatment interactions where there are none.

This chapter will begin with a discussion of the quantification of the treatment effect. We will then present hypothetical across-study data showing the effects of sampling error and the failure of the conventional statistical significance test in the context of the review study. We will then present a substantive discussion of artifacts other than sampling error. These other artifacts can be just as large in size even though they are usually systematic rather than random in nature.

Quantification of the Treatment Effect:
The $d$ Statistic and the Point Biserial Correlation

A key issue is the description of treatment effects as quantitative or dichotomous. The traditional description is dichotomous: The treatment either had an effect or had no effect. Methodologists have long argued that we should instead describe the treatment effect in quantitative form, that is, estimate the actual size of the treatment. A dichotomous description is poor for several reasons. First, there is a great loss of information, information that can be used (1) to assess the practical importance of a treatment, (2) to compare the effectiveness of treatments, (3) to determine whether a theory has been confirmed or disconfirmed, and (4) to test quantitative theories such as path models. Second, the implicit assumption in dichotomizing the treatment effect is that most treatments have no effect. If this were true, then there would be important information in the statement that the treatment effect is not 0. However, as discussed previously, meta-analyses have now shown that treatments rarely have no effect at all. The conclusion, “The treatment had no effect,” is usually erroneous. Thus, the question for a treatment is really not whether it had an effect, but whether the effect is as large as a theory predicts, whether the effect is large enough to be of practical importance, or whether the effect is larger or smaller than some other treatment or some variation of the treatment. These questions can only be answered by quantifying the size of the treatment effect.

The dichotomization of treatment effects is also related to the statistical analysis of treatments. If it were true that most treatments have no effect, then good statistical analysis would focus on Type I error: falsely concluding that there is an effect when there is no such effect. The conventional significance test guarantees that
Type I errors will occur no more than 5% of the time. However, meta-analysis has now shown that this nihilistic null hypothesis is rarely true. If the null hypothesis is false, then all statistical errors will be Type II errors: falsely concluding that there is no effect when there is, in fact, an effect. As we shall see, for typical sample sizes, the Type II error rate is quite high. For sample sizes of 100, the Type II error rate for textbook experiments is around 50% and the Type II error rate for more subtle follow-up research is higher yet. There are many important research domains where the significance test error rate is as high as 85%.

Because the null hypothesis is false in most research domains, the conventional significance test has a very high error rate. This high error rate means that the conventional significance test is actually counterproductive at the level of review studies. The high error rate for the conventional significance test means that results interpreted using the significance test must necessarily look inconsistent across studies. For example, if the significance test is wrong 50% of the time, then half the studies will have a significant treatment effect, but the other half will falsely appear to show no treatment effect.

This is quite evident in comparing the results of meta-analyses to the conclusions of narrative reviews. For most questions studied, meta-analysis shows that the treatment effect was not 0—although treatment effects are sometimes quite small. Narrative reviews, on the other hand, have been inconsistent. Some reviewers are selective; they throw out studies on “methodological” grounds—frequently of an entirely hypothetical nature. They throw out studies until those that remain have consistent results. They then base their conclusions on the remaining studies. Unfortunately, different reviewers will throw out different studies and, hence, come to different—sometimes opposite—conclusions. Comprehensive reviewers make a different error: They usually conclude that treatment effects are sporadic. They conclude that the treatment effect is present in some studies but absent in others.

The natural quantitative description of the treatment effect is just the difference between the means on the dependent variable. Let $Y$ be the dependent variable. Denote the means for the control and experimental groups as follows:

- $\bar{Y}_E =$ the mean for the experimental group
- $\bar{Y}_C =$ the mean for the control group

To say, “The treatment increased performance by 3.2 feet,” is to say that the difference $\bar{Y}_E - \bar{Y}_C$ is 3.2 feet, that is,

$$\bar{Y}_E - \bar{Y}_C = 3.2$$

If the dependent variable were identically measured in all studies, then the raw score difference between means would be the conventional measure of the treatment effect. However, this is rarely true. Consider the measurement of the job performance of sewing machine operators. One would think that a measure such as “number of garments sewn per week” would be the same variable across studies. However, workers at different places are sewing different kinds of garments. To sew three dresses might be very different from sewing three coats. Thus, typically, the units of the dependent variable vary from one study to the next.
If the dependent variable is the same in two different studies except for units, then it would, in principle, be possible to calibrate the two measures by finding the constant of proportionality between the two units. Consider, however, the problem of matching the units for sewing machine operators in two different studies. In one study, the workers sew dresses, while the workers in the other study sew coats. To transform scores from one metric to the other, the workers at one place would have to sew the other kind of garment. Furthermore, they would have to be given exactly the same training in sewing that other kind of garment to be exactly comparable. This would be prohibitively expensive even if it were possible. Thus, exact calibration of independent variables is also impossible in most research domains.

There is an alternative method of matching across studies, although it depends on a substantive assumption. We can eliminate units within a study by using standard scores instead of raw scores. The treatment effect in standard scores would then be given by

\[ d = \frac{\overline{Y}_E - \overline{Y}_C}{\sigma} \]

where \( \sigma \) is the standard deviation of the raw scores in that study. The only question is, “Which standard deviation?” This question will be considered in detail in the next chapter. For population data, the natural definition would be to use the population standard deviation of the control group. However, for sample data, the standard deviation is much better estimated by using the “within-group variance,” that is, by averaging the experimental and control group standard deviations. This sample statistic is Cohen’s (1977) “\( d \) statistic,” which is the most widely used statistic in the meta-analysis of experimental or intervention studies. For the population value, we will use the Greek letter for \( d \), that is, \( \delta \).

Suppose the distribution of garment sewing performance per month has a mean of 100 and a standard deviation of 25. If a training program increases performance by 10 garments per day, then the treatment effect in standard scores would be

\[ d = \frac{10}{25} = .40 \]

That is, the treatment effect would be .40 standard deviations.

If the outcome (dependent) variable is a true dichotomy (e.g., patient had the disease vs. patient did not have the disease), then another statistic, the odds ratio, can be used. The odds ratio is frequently used in medical research. We do not present procedures for using the odds ratio in this book because it is rarely appropriate and is rarely used in social science research. Haddock, Rindskopf, and Shadish (1998) provided a discussion of potential uses of the odds ratio in social science research.

There is a closely related measure of treatment effect that will be discussed in detail in the next chapter: the point biserial correlation. The point biserial correlation is actually an ordinary Pearson correlation; the special name comes from the nature of the data on which it is computed. We create a single data set by pooling the data across the control group and the experimental group. We define a treatment variable (sometimes called a “dummy variable” or “contrast
variable”) by assigning different scores to the people in the two different groups. For example, we might define the variable T by assigning the score 0 to those in the control group and assigning the score 1 to those in the experimental group. The correlation computed on the pooled data between that treatment variable and the dependent variable is the point biserial correlation. The point biserial correlation has the advantage that it can be treated like any other correlation coefficient. In particular, the meta-analysis could be done using the methods of Chapters 3 and 4 on the correlation coefficient. The mathematics is then much easier than that for the d statistic. The correlation is much easier to fit into advanced statistical analyses such as reliability analysis, path analysis, and so on. The point biserial correlation is the second most often used quantification of the treatment effect in meta-analysis. As noted in the next chapter, the two statistics, r and d, can be algebraically transformed back and forth from each other. Thus, it is conceptually arbitrary which statistic is used. However, in this chapter, we will primarily use d.

For the usual empirical range of d of \(-.41 < d < +.41\), the conversion formulas between r and d are trivial.

\[
\begin{align*}
  d &= 2r & \text{for } -0.21 < r < +0.21 \\
  r &= 0.5d & \text{for } -0.41 < d < +0.41
\end{align*}
\]

How close is this approximation? Consider the worst case, \(d = 0.40\). The approximation \(0.5d\) yields \(r = 0.20\), while the actual correlation is 0.196.

The d statistic is comparable across studies if the standard deviation of the dependent variable (measured in any one set of units) is the same across studies. This is a typical finding for standardized variables in psychology in the absence of processes producing range restriction. Although means often differ considerably from one setting to the next, standard deviations often differ little. In a research domain where this is not true, the variation in results due to differing units could only be corrected by making a “correction for range variation” (see Chapter 3).

### Sampling Error in d Values: Illustrations

Is an argument more effective if it is expressed in intense language or if it is cast in wishy-washy language? A meta-analysis by Hamilton and Hunter (1987) showed the difference in attitude change to be about 0.20 standard deviations (i.e., \(d = 0.20\) or \(r = 0.10\)) favoring strong language. Assume that this is the population value of the d statistic for all studies in a hypothetical meta-analysis. What would the review data look like? That depends on the sample sizes used in the studies collected. For simplicity, suppose all studies had used exactly the same sample size. The study results would be approximately distributed as in Table 6.1. (Note: The distributions in Table 6.1 exactly match the sampling distribution for replicated studies. An actual 19-study meta-analysis would find values that departed from this distribution somewhat because the 19 observed sampling errors would not match the exact population distribution of sampling errors.)
Table 6.1  Hypothetical meta-analysis data for the effect of language intensity on persuasion (results ordered by magnitude)

<table>
<thead>
<tr>
<th>Study</th>
<th>N = 30</th>
<th>N = 68</th>
<th>N = 400</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.80**</td>
<td>0.60**</td>
<td>0.36**</td>
</tr>
<tr>
<td>2</td>
<td>0.68*</td>
<td>0.50**</td>
<td>0.32**</td>
</tr>
<tr>
<td>3</td>
<td>0.58</td>
<td>0.46*</td>
<td>0.30**</td>
</tr>
<tr>
<td>4</td>
<td>0.50</td>
<td>0.40*</td>
<td>0.28**</td>
</tr>
<tr>
<td>5</td>
<td>0.44</td>
<td>0.36</td>
<td>0.26**</td>
</tr>
<tr>
<td>6</td>
<td>0.40</td>
<td>0.32</td>
<td>0.26**</td>
</tr>
<tr>
<td>7</td>
<td>0.34</td>
<td>0.30</td>
<td>0.24**</td>
</tr>
<tr>
<td>8</td>
<td>0.30</td>
<td>0.26</td>
<td>0.22**</td>
</tr>
<tr>
<td>9</td>
<td>0.24</td>
<td>0.24</td>
<td>0.22**</td>
</tr>
<tr>
<td>10</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20**</td>
</tr>
<tr>
<td>11</td>
<td>0.16</td>
<td>0.16</td>
<td>0.18*</td>
</tr>
<tr>
<td>12</td>
<td>0.10</td>
<td>0.14</td>
<td>0.18*</td>
</tr>
<tr>
<td>13</td>
<td>0.06</td>
<td>0.10</td>
<td>0.16</td>
</tr>
<tr>
<td>14</td>
<td>-0.00</td>
<td>0.08</td>
<td>0.16</td>
</tr>
<tr>
<td>15</td>
<td>-0.04</td>
<td>0.04</td>
<td>0.14</td>
</tr>
<tr>
<td>16</td>
<td>-0.10</td>
<td>0.00</td>
<td>0.12</td>
</tr>
<tr>
<td>17</td>
<td>-0.18</td>
<td>-0.06</td>
<td>0.10</td>
</tr>
<tr>
<td>18</td>
<td>-0.28</td>
<td>-0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>19</td>
<td>-0.40</td>
<td>-0.20</td>
<td>0.04</td>
</tr>
</tbody>
</table>

**Significant by two-tailed test.
*Significant by one-tailed test.

NOTE: In each case, the population effect is given by $\delta = 0.20$ in all studies and all deviation from that value is entirely due to sampling error. The sample size is the total sample size across control (low intensity) and experimental (high intensity) groups. Thus, “N = 30” means “15 in each group.”

**Case 1: N = 30**

Suppose 19 studies were done with a total sample size of 30 (15 subjects in each group) in each study. The study treatment effects would distribute themselves as the first column of Table 6.1. Six of the studies would have had negative observed treatment effects. The authors of these studies would believe that intense language is counterproductive and reduces the persuasive effect. On the other hand, six studies would have found treatment effects of $d = 0.40$ or more, effects as large as textbook examples. These authors would believe that intense language is one of the most powerful persuasive agents known. Both sets of authors would be wrong. Only Study 10—the median study—has an effect size of $d = 0.20$, the actual population value for all studies.

One classic but crude method of reviewing research is to count the number of studies in the predicted direction. This count is 13 out of 19. This is greater than the 9.5 out of 19 expected by chance, though not significantly so (using a binomial test). However, had there been 190 studies instead of 19, the expected count of studies in the predicted direction would be 130/190, which is significantly greater than the 95 expected by chance. Thus, a count of the studies in the predicted direction would show that intensity increased persuasion more often than chance. However, it would falsely suggest that intensity acted in the *opposite* direction 32% of the time.
The statistical significance test was designed to reduce the impact of sampling error. When the null hypothesis is true, it should reduce errors of inference to 5%. How does the conventional significance test fare in this example? There are two ways to do this significance test. Had each study been analyzed using analysis of variance, it would have been analyzed using a two-tailed significance test, and only the study with \( d = .80 \) would have been significant. That is, analysis of variance yields the correct inference for only one study, an error rate of 18/19, or 95%. That is, the error rate for the two-tailed significance test is not 5% but 95% in this example.

Had the data been analyzed using the \( t \) test, the authors would have had the option of doing a one-tailed test. For a one-tailed test, both Studies 1 and 2 have significant treatment effects. Thus, this significance test yields a correct inference for only 2 of the 19 studies, an error rate of 17/19, or 89%. Thus, for a one-tailed \( t \) test, the error rate is not 5% but 89% in this example.

In this example, the two-tailed test (conventional analysis of variance) is correct in only 1 of 19 studies. The one-tailed test is correct in 2 of 19 studies, which doubles the power of the two-tailed test. However, in either case the error rate is far higher than the 5% error rate that most people believe to be the error rate for the statistical significance test.

Why is the error rate higher than 5%? The conventional statistical significance test assumes a nihilistic null hypothesis of \( \delta = 0 \). If the null hypothesis were true, then the error rate would be only 5%. However, the null hypothesis is false for this research domain (as it is in most research domains), and thus, the error rate is not constrained to be 5%, but will be higher. In this example, the error rate rose to 89% (one-tailed test) or 95% (two-tailed test), which is close to the theoretical maximum error rate.

Consider the position of a reviewer faced with study results such as those in Table 6.1. If the reviewer counts results in the expected direction, then there is a weak indication of results in the expected direction. It is true that nearly a third of the studies go in the wrong direction, but that is counterbalanced by the third of the studies with effects as large as classic textbook effects in social psychology. That reviewer would probably conclude that intense language is more persuasive most of the time but would warn that there are some settings where, for unknown reasons, intense language is counterproductive. This would be a false interpretation of the data.

Suppose the reviewer ignored the size of the treatment effects and considered only a count of the number of significant findings using a two-tailed test. This reviewer would almost certainly conclude that language intensity has no effect on persuasiveness. That, too, would be a false conclusion. Ironically, the reviewer who uses the significance test—the more “sophisticated” method—is even farther off base than the reviewer who naively looks at raw results!

Note that the inferences of reviewers would not materially improve with more data. If the number of studies rose from 19 to 190, the number of studies with results significant by a one-tailed test would rise from 2 to 20. However, the proportion of significant findings would still be the same, 20/190 = 2/19. Thus, a reviewer who depended on the significance test would still draw the same false conclusions even though there were 10 times as much data.
As we will see, the method of meta-analysis presented in this book will deal with these data correctly. This method would estimate the average treatment effect to be $\delta = .20$ to within the sampling error left by using a total sample size of $N = 19(30) = 570$. If there were 190 studies, the error in estimating the mean effect size would drop to that left by a total sample size of $\bar{N} = 190(30) = 5,700$. As more and more studies become available, this method of meta-analysis has less and less error. This method would also have correctly concluded that all or nearly all the variance in observed study effects was due to sampling error.

Is this example far-fetched? The size of the treatment effect for language intensity is that found in actual studies. On the other hand, the sample size of $N = 30$ is lower than the actual studies ($\bar{N} = 56$). However, there are important research domains with sample sizes this low. For example, Allen, Hunter, and Donahue (1988) did a meta-analysis on studies of the effect of psychotherapy on problems of shyness and fear of public speaking. For the studies using systematic desensitization, the average sample size was 23. For the studies using rational-emotive therapy, the average sample size was only 19.

**Case 2: $N = 68$**

The median sample size of studies in personnel selection is 68 in the pre-1980 literature (Lent, Auerbach, & Levin, 1971a, 1971b). This seems not far from sample sizes in other psychological study domains although there are exceptions both larger and smaller. The average sample size for the language intensity meta-analysis done by Hamilton and Hunter (1987) was $\bar{N} = 56$, which is about the same as the 68 used in Table 6.1. If all 19 studies were done with a sample size of $N = 68$, then the study values would have an expected distribution like that of the second column of Table 6.1.

A reviewer who looked at the results at face value would now see 15 of 19 values in the expected direction and only 4 of 19 negative values. This split is significantly different from a 50–50 split using a binomial comparison. At the same time, the four large values are not quite as large as textbook examples. This reviewer would probably conclude that the studies in the wrong direction were just sampling errors from a zero effect. Thus, the reviewer would probably conclude that language intensity usually increases persuasion, although there are a minority of cases where it does not. This conclusion is false because the effect is actually $\delta = .20$ in all cases.

The conventional two-tailed statistical significance test of analysis of variance registers only the two largest values as significant. Thus, the conventional two-tailed test is correct in only 2 of 19 cases, an error rate of 17/19, or 89%. A reviewer who counted significant findings would probably conclude that language intensity is irrelevant to persuasion. This conclusion would be a grave error in this example.

The one-tailed significance test registers the top four values as significant. Thus, the one-tailed test is correct 4 times, which means that the one-tailed test has twice the power of the two-tailed test in this example. However, the one-tailed test is still wrong in 15 of 19 studies, an error rate of 79%. A reviewer who counts one-tailed significant findings would probably conclude that 4 times in 19 is noticeably
greater than the 1 in 20 expected by chance. If not, then if the number of studies were raised to 190, the reviewer would certainly notice that 40 out of 190 is much greater than the $190/20 = 9.5$ expected by chance. The reviewer would probably conclude that language intensity does have an impact in about $(40 - 10)/190$, or 16%, of settings, but has no effect otherwise. This is an improvement over the error made by the reviewer who looks at two-tailed tests, but is worse than the conclusion drawn by the reviewer who ignores the significance test altogether.

The method of meta-analysis presented here would estimate the treatment effect to within the sampling error left by a total sample size of $N = 19(68) = 1,292$. If there were 190 studies, the error in the mean effect size would be down to that left by a total sample size of $N = 190(68) = 12,920$. The method would also correctly conclude that all or nearly all of the variance across studies is due to sampling error.

Case 3: $N = 400$

Most psychologists think of a sample size of 400 as if it were $\infty$. However, pollsters know differently from experience. The typical study results for 19 studies with a sample size of $N = 400$ are shown in the third column of Table 6.1.

A reviewer who looks at the results at face value would now note that all results are in the expected direction, although the smallest results are small indeed. The largest results are still moderate in size. Thus, the reviewer would probably conclude that language intensity always increases persuasion (a correct conclusion) although in some settings the impact is negligible in magnitude (an incorrect conclusion).

A reviewer who counts two-tailed significance tests would find that 10 of 19 study values are significant. This reviewer would probably conclude that language intensity increases persuasion in about half of the settings but does not work in the other half. This conclusion is quite far from the truth.

A reviewer who counts one-tailed significance tests would find that 13 of 19 study values are significant. Thus, in this example, the one-tailed test is $13/10$ times more powerful than the two-tailed test, that is, about $30\%$ more powerful. This reviewer would probably conclude that language intensity increases persuasion in about two-thirds of the settings, but does not work in the other third. This conclusion is also quite far from the truth.

Even with a sample size of 400, the reviewer who naively looks at face value results is closer to the truth than a reviewer who counts statistical significance findings. Thus, even with a sample size of 400, the significance test still works so poorly that it is counterproductive in comparison to doing no analysis for sampling error at all.

With an average sample size of 400, our method of meta-analysis would estimate the mean effect size to within the sampling error left by a total sample size of $N = 19(400) = 7,600$. The analysis would also correctly conclude that all or nearly all of the variance across studies is due to sampling error.

From the viewpoint of review studies, the statistical significance test does not correctly deal with sampling error. The statistical significance test works only in a
research context in which we know the null hypothesis to be true. If we know the null hypothesis to be true, however, then we need not do the test at all. Thus, we should abandon the use of the statistical significance test in doing review studies. There are now many sets of mathematically equivalent meta-analysis formulas that take sampling error into account correctly for mean effect sizes, including the method presented here. Our method will also work when there is real variance in effect sizes across studies. We will estimate the size of the standard deviation of population effect sizes. Some authors stop with a significance test for homogeneity and present no method for estimating the standard deviation if the significance test indicates that the standard deviation is not 0.

Error of Measurement in the Dependent Variable

Ordinary English interprets the phrase “error of measurement” as having two meanings: systematic and unsystematic error. Systematic error is a departure from measuring exactly what was intended. In psychometric theory, this is called “imperfect construct validity.” In psychometric theory, the phrase “error of measurement” is used for unsystematic error, also called “random error” or “unreliability.” We will follow psychometric terminology here. This section will present the effects of unsystematic or random error of measurement and a later section will cover imperfect construct validity.

In psychology, much of the unsystematic error of measurement is caused by randomness in subject response. This kind of error usually has a mean of 0, that is, equally likely to be positive or negative, and is uncorrelated with the true value. If we write the observed score on the dependent variable as \( Y \), write the true score as \( U \), and write the error of measurement as \( e \), then

\[
Y = U + e
\]

where the population mean of \( e \) is 0 and the population correlation between \( e \) and \( U \) is 0.

Because the average error is 0, the mean of errors does not describe the typical size of an error. Rather, the typical size of errors is described by either the error variance—the average squared error—or the error standard deviation. The number \( \sigma_e \) is called the “standard error of measurement” in psychometric theory. The practical impact of error of measurement is relative to the size of differences between people. If two people differ on the dependent variable by 10 points, then errors of size \(-1\) or \(+1\) would have little effect on the comparison of those people. On the other hand, if the difference between two subjects were .5, then errors of \(-1\) or \(+1\) would completely obscure the comparison. One measure of the relative error of measurement is the “noise to signal” ratio, \( \sigma_e / \sigma_U \), although this is not commonly used. Instead, the more useful measure of relative error is the correlation between true and observed score, that is, \( r_{UY} \). By historical convention, the square of this correlation is called the “reliability” of the dependent variable and is denoted \( r_{UY}^2 \). That is, we define the reliability of the dependent variable \( r_{YY} \) by

\[
r_{YY} = r_{UY}^2
\]
Different ways of estimating reliability identify and assess different sources of measurement error. It is critical that the researcher use the appropriate reliability estimate. We refer the reader to the extended treatment of this issue presented in Chapter 3. The error standard deviation and the reliability of the dependent variable are related by

$$\sigma_e = \sigma_Y \sqrt{1 - r_{YY}}$$

The size of the reliability depends on the extent of measurement error in the process measured—usually a response in psychology—and on the number of primary measurements used to generate the final response—frequently the number of items on a scale. High-quality measurement often provides reliability in the region of \(r_{YY} = .81\). Moderate quality usually falls around \(r_{YY} = .64\). Measurement based on a single response frequently has reliability no higher than \(r_{YY} = .25\). It should be noted that the reliability of a single response is not determined by the cost of obtaining that response. For example, in equity studies in social psychology, subjects may spend as much as an hour before the criterion act. However, the only measurement of the dependent variable is a single response: the amount of money given to the partner. The reliability of that single response is the correlation between that response and the response that would have been made on some other randomly chosen day. The reliability of single responses is rarely higher than \(r_{YY} = .25\).

The size of the reliability depends both on the extent of error in the measurement process and on the extent of individual differences on the dependent variable. For instance, Nicol and Hunter (1973) found that the same semantic differential scale that had a reliability of .90 measuring attitudes toward the polarized issue “law and order” had only a reliability of .20 measuring attitudes toward the issue “pollution.”

The observed score for a given person \(p\) is related to the true score for that person by

$$Y_p = T_p + e_p$$

If we average scores across persons, the mean score is related to the mean true score by

$$\bar{Y} = \bar{T} + \bar{e}$$

That is, errors of measurement are averaged across persons. The population mean of scores across persons averages the errors of measurement across an \(\infty\) (infinity) of errors and is thus 0. That is, at the population level, error of measurement has no impact on the mean.

The raw score treatment effect is defined as the difference between population means:

$$\text{Raw score } \delta_Y = \bar{Y}_E - \bar{Y}_C$$

Because population mean error of measurement is 0, each mean observed score is equal to the mean true score. Thus,

$$\text{Raw score } \delta_U = U_E - U_C = \bar{Y}_E - \bar{Y}_C = \text{Raw score } \delta_Y$$
That is, random error does not alter the raw score treatment effect. This is the reason that traditional statistics has ignored error of measurement in the treatment of experimental design.

However, it is not the raw score treatment effect but rather the standard score treatment effect that is of primary interest in statistics. For purposes of meta-analysis, it is normally necessary to use standard score treatment effects to achieve comparability across studies. However, the standard score treatment effect is also central to traditional statistics because it is the standard score treatment effect that is assessed by the statistical test for significance. In particular, the power of the conventional significance test depends on the standard score treatment effect.

Error of measurement does not affect the mean of the dependent variable but it does affect the variance. The variance of observed scores is related to the variance of true scores by

$$\sigma_Y^2 = \sigma_U^2 + \sigma_e^2$$

That is, error of measurement increases the variance and, hence, the standard deviation of the dependent variable. Consider, then, the experimental versus control group comparison. Adding error does not change the means, but it increases the spread of scores about the mean. This effect is shown in Figure 6.1.

The extent of separation between two groups depends on the extent of overlap between the two distributions. The extent of overlap between the distributions depends on the difference between the means in relation to the extent of spread about the means. The greater the spread about the means, the greater the overlap between the two distributions. Figure 6.1 shows that the extent of overlap is greatly increased by the presence of error of measurement. The lower the reliability, the larger the spread about the means and, hence, the greater the overlap. That is, as the amount of error of measurement increases, the difference is more and more obscure. In terms of statistical power, the more obscure the difference between the means, the more difficult that difference is to detect.

Consider the standardized effect size for true scores and observed scores:

$$\delta_U = (U_E - U_C)/\sigma_U$$
$$\delta_Y = (Y_E - Y_C)/\sigma_Y$$

Because population means are not affected by error of measurement, the numerators are equal. However, error of measurement increases the standard deviation and, hence, the denominators are different. The increase in standard deviation is given by

$$\sigma_Y = \sigma_U / \sqrt{r_{YY}}$$

where we note that to divide by a number less than 1 is to increase the ratio. If this identity is substituted into the equation for $\delta_Y$, we have

$$\delta_Y = \delta_U / \sqrt{r_{YY}}$$

That is, the standardized effect size for the observed score is the standardized effect size for the true score multiplied by the square root of the reliability. For example, if the reliability were $r_{YY} = .81$, the effect size would be reduced to

$$\delta_Y = .90 \delta_U$$

that is, reduced by 10%.
If the effect size is reduced by error of measurement, the effect is more difficult to detect by the conventional significance test. This is illustrated in Table 6.2. Table 6.2 computes the power of the conventional significance test for studies of sample size $N = 100$, a value roughly typical for empirical studies. An effect size of $\delta = .40$ is about the size of large introductory textbook examples from the social psychology experimental literature. The effect size of $\delta = .20$ is about the size of effects in the more sophisticated research that follows up textbook examples, that is, research that studies variation in textbook manipulations rather than the crude manipulation itself.

Table 6.2 first shows the reduction in the treatment effect produced by different levels of error of measurement. As the reliability decreases from $r_{YY} = 1.00$ to $r_{YY} = .25$, the treatment effect is reduced by half, for example, from $\delta = .40$ to $\delta = .20$ or from $\delta = .20$ to $\delta = .10$. The probability of detecting the effect using a conventional significance test drops correspondingly. For textbook size effects with $\delta = .40$, the power drops from an already low 51% to 17%, a power level
Table 6.2  Power of the conventional significance test for studies with sample size \(N = 100\)

I. Reduction in the effect size: the value of \(\delta_y\) for various values of the reliability of the dependent variable

<table>
<thead>
<tr>
<th>Reliability</th>
<th>(\delta_U = .10)</th>
<th>(\delta_U = .20)</th>
<th>(\delta_U = .30)</th>
<th>(\delta_U = .40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>.10</td>
<td>.20</td>
<td>.30</td>
<td>.40</td>
</tr>
<tr>
<td>.81</td>
<td>.09</td>
<td>.18</td>
<td>.27</td>
<td>.36</td>
</tr>
<tr>
<td>.64</td>
<td>.08</td>
<td>.16</td>
<td>.24</td>
<td>.32</td>
</tr>
<tr>
<td>.25</td>
<td>.05</td>
<td>.10</td>
<td>.15</td>
<td>.20</td>
</tr>
</tbody>
</table>

II. Reduction in power: the power of the conventional .05 level statistical significance test for various values of the reliability of the dependent variable, expressed as a percentage

<table>
<thead>
<tr>
<th>Reliability</th>
<th>(\delta_U = .10)</th>
<th>(\delta_U = .20)</th>
<th>(\delta_U = .30)</th>
<th>(\delta_U = .40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>7.2</td>
<td>16.6</td>
<td>31.8</td>
<td>51.2</td>
</tr>
<tr>
<td>.81</td>
<td>6.5</td>
<td>14.3</td>
<td>26.7</td>
<td>43.2</td>
</tr>
<tr>
<td>.64</td>
<td>5.9</td>
<td>12.1</td>
<td>33.0</td>
<td>37.3</td>
</tr>
<tr>
<td>.25</td>
<td>4.3</td>
<td>7.2</td>
<td>11.2</td>
<td>16.6</td>
</tr>
</tbody>
</table>

only one-third the size. Stated the other way, the error rate for the significance test rises from 49% to 83%. For sophisticated research, the initial power level is smaller to begin with and, hence, there is less distance to fall. If with perfect measurement the effect size is \(\delta = .20\), then, if the dependent variable is based on a single response, the observed effect size would be about \(\delta = .10\) and the power would drop from 17% to 7%, a reduction of slightly more than half. The error rate for the significance test rises from 83% to 93%.

The next chapter will show that for population effect sizes the reduction in the treatment effect can be corrected if the reliability of the dependent variable is known. Although the application of that same correction formula to sample values eliminates the systematic effect of error of measurement (Cook et al., 1992, pp. 315–316), the increased sampling error and reduced statistical power cannot be corrected. This shows in the fact that the significance test on the corrected effect size is algebraically equivalent to the significance test on the uncorrected effect size.

Error of Measurement in the Treatment Variable

The treatment variable is defined by group assignment. Because the investigator usually knows exactly which group each subject belongs to, this variable is usually regarded as perfectly measured. However, this definition ignores the interpretation of the results. In interpreting the results, it is not group assignment but the treatment process that is assumed to be the independent variable. From that point of view, the nominal treatment variable may be quite different from the actual treatment variable.

Consider an attitude change experiment directed to the topic of acid rain. The investigator seeks to manipulate the credibility of the source of the change-inducing message. At one point in the instructions, the sentence “The author of the message is . . .” is completed either with the phrase “a famous scientist in the area” or
with “a marine drill instructor.” Assume that the instructions are correctly read to each subject so that we do know correctly and exactly the group assignment of each subject. However, suppose 30% of the subjects are not paying careful attention to the instructions. They do not hear the sentence stating the author of the message. They do not think about the author of the message until they read it. Assume that the subjects who did not hear the author’s identity then assume one. Suppose half the subjects assume the author to be an expert while half the subjects assume the author to be some know-nothing graduate assistant. Then, in this study, 15% of the control group subjects will assume an expert source and will act as if they had been exposed to the experimental group instructions, while 15% of the experimental group subjects will assume a know-nothing source and will act as if they had been exposed to the control group instruction. In this example, the nominal treatment group variable is the reverse of the actual treatment variable 15% of the time. The observed effect size will be correspondingly reduced.

How much will the treatment effect be reduced? The idea is simple, but the computations are complicated. In our example, where 15% of the subjects were misidentified in each group, we could compute the effect of the treatment variable error by assuming each treatment group to be the pooling of 85% from the corresponding true treatment group and 15% from the other true treatment group. The outcome is easily stated in correlational terms. Denote the nominal treatment variable by \( X \) and denote the true treatment variable by \( T \), that is,

\[
X = \text{the observed group assignment of the subject},
\]

\[
T = \text{the actual treatment value for that subject}.
\]

If the correlation between \( X \) and \( T \) is \( r_{YT} \), then the observed treatment effect correlation \( r_{XY} \) is related to the true treatment effect correlation \( r_{TY} \) by the equation

\[
r_{XY} = r_{XT} r_{TY}
\]

The formula for the reduction in the \( d \) statistic can be obtained by substituting this product into the formula for \( r \) to \( d \) conversion.

The product rule for the treatment effect correlation is a special case of the attenuation formula from psychometric theory. Let us denote the “reliability of the treatment” by \( r_{XX} \) and define it to be the square of the correlation between true and observed treatment identifications; that is, we define \( r_{XX} \) by

\[
r_{XX} = r_{XT}^2
\]  

(6.1)

Then our product formula is a special case of the psychometric formula

\[
r_{XY} = \sqrt{r_{XX} r_{TY}}
\]  

(6.2)

In our attitude change example, we assumed 15% misidentification in each group. The correlation between the observed and true treatment is thus \( r_{XT} = .70 \), and, hence, the observed treatment effect correlation is

\[
r_{XY} = .70 r_{TY}
\]

That is, the observed treatment effect correlation is reduced by 30%.
If the treatment effect correlation is reduced, then statistical power will be reduced correspondingly. Suppose in our example that the true treatment effect was $\delta_T = .40$ with a sample size of $N = 100$. The true treatment effect correlation would then be $r_{TY} = .20$. The observed treatment effect correlation would then be $r_{XY} = (.70)(.20) = .14$ and the observed treatment effect would be $\delta_X = .28$. Had there been no error in the treatment identification, the statistical power would have been 51%. Instead, it is 28%, reduced by nearly half.

If the reliability of the treatment variable is known, then the attenuation effect can be corrected. The formula is the usual psychometric formula for correction for attenuation due to error in the independent variable:

$$r_{TY} = r_{XY} / \sqrt{r_{XX}} \quad (6.3)$$

This correction works perfectly at the population correlation level. However, the correction at the sample data level corrects for only the systematic attenuation. It does not correct for the increased sampling error introduced by the measurement error. The significance test on the corrected correlation is algebraically equivalent to the significance test on the uncorrected correlation.

In the previous example of attitude change experiment, errors of measurement in the independent variable had two effects: (1) Within both the experimental and control groups, the within-group variance on the dependent variable was increased; and (2) the raw score mean difference on the dependent variable was reduced, that is, $\bar{Y}_E - \bar{Y}_C$ was reduced. In such a case, the observed $d$ value will be reduced for two reasons: because the numerator is reduced and because the denominator is increased. There are other cases of measurement error in the independent variable in which the numerator, $\bar{Y}_E - \bar{Y}_C$, is unaffected but the denominator, the pooled within-group SD, is inflated, leading to artifically lowered estimates of $d$.

For example, suppose an experiment is conducted to determine the effect of personal attention and sympathetic listening by work counselors on the job-related attitudes of problem employees. Each member of the experimental group is supposed to get 12 hours of personal interaction (6 two-hour sessions) with a counselor. However, because of interruptions of scheduled sessions, lateness, and other problems, some people in each study get less than that: some 10 and some 11 hours. Because some counselors run past the stopping time without realizing it, other members of the experimental group get more than 12 hours: some 13 and some 14 hours. The average amount of time might be approximately correct: 12 hours. If the impact of treatment strength differences is approximately linear over the range of variation in the study (true in most cases), then the average effect will be determined by the average treatment strength. The individual variations will cancel out, and the mean of the treatment group will be the same as if there had been no variation in treatment. Thus, the numerator of the effect size formula for $d$ (i.e., $\bar{Y}_E - \bar{Y}_C$) will not be affected. However, the individual variations in treatment strength will cause variations in outcome that will contribute to variation in the dependent variable. Thus, the denominator of the effect size will be larger than would be true if there were no variation in treatment strength. If the denominator of the effect size were increased, then the effect size would be reduced. Thus, within-study variation in treatment strength that has no effect on $\bar{Y}_E - \bar{Y}_C$ nevertheless reduces the effect size.
Furthermore, because the extent of within-study variation is likely to differ from one study to the next, failure to correct for attenuation due to treatment variation will lead to artificial variation in effect size across studies. This uncorrected variation could be falsely interpreted as showing the existence of a nonexistent moderator variable.

Variation in the treatment effect increases the experimental group standard deviation, but does not change the control group standard deviation. The increase in the experimental group standard deviation increases the within-group standard deviation and, hence, reduces the observed effect size value. However, this artificial increase in the experimental standard deviation could also cause another error. If there were no true treatment by subject interaction and if there were no variation in the treatment effect, then the control and experimental group standard deviations would be equal. The artificial increase in the experimental group standard deviation might be falsely interpreted as an indication of a treatment by subject interaction. That is, it appears that subjects in the experimental group are reacting differently to the same treatment (causing $SD_E$ to be larger than $SD_C$), when, in fact, the cause of the larger $SD_E$ is the fact that different subjects in the experimental group are by mistake receiving treatments of different intensity or duration.

If there is no true treatment by subject interaction, then the increase in the experimental group standard deviation can be used to quantify the impact of treatment variation. If there is no interaction, then the desired effect size is

$$\delta = (\bar{Y}_E - \bar{Y}_C)/SD_C \quad (6.4)$$

The observed population effect size is

$$\delta_o = (\bar{Y}_E - \bar{Y}_C)/SD_W \quad (6.5)$$

The two effect sizes differ by

$$\delta_o = a\delta \quad (6.6)$$

where the attenuation factor $a$ is given by

$$a = SD_C/SD_W \quad (6.7)$$

For equal sample sizes, the attenuation factor can be computed from the ratio comparing the experimental and control group standard deviations. Denote the standard deviation comparison ratio by $v$. That is, define $v$ by

$$v = SD_E/SD_C \quad (6.8)$$

Then the within-group standard deviation is related to $v$ by

$$SD_W = \sqrt{[(SD_C^2 + SD_E^2)/2]} = SD_C\sqrt{[(1 + v^2)/2]} \quad (6.9)$$

Thus,

$$a = SD_C/SD_W = 1/\sqrt{[(1 + v^2)/2]} \quad (6.10)$$
If $v$ is not much larger than 1, then we have the approximation

$$a = 1 - \frac{(v^2 - 1)}{2} \quad (6.11)$$

In summary, within-study variation in treatment strength causes an inflation in the experimental dependent variable standard deviation. If there is no real treatment by subject interaction, then variation in treatment strength causes the experimental group standard deviation to be artificially larger than the control group standard deviation. If treatment variation is not suspected, then this increase could be falsely interpreted as indicating a treatment by subject interaction. If it is known that there is no interaction, then the attenuation in the effect size can be computed from a comparison ratio of the experimental to control group standard deviation.

**Variation Across Studies in Treatment Strength**

In the preceding example, the mean raw score treatment effect, $\bar{Y}_E - \bar{Y}_C$, is the same in all studies. In other cases, however, this value may vary across the studies in a meta-analysis—because the amount of treatment given to the experimental group might differ, causing $\bar{Y}_E$ to vary across studies. If these differences are known (i.e., if they are given in each study), they can be coded and treated as a potential moderator variable. If the strength of treatment values is not known, however, then variation in treatment strength will produce variation in effect sizes that cannot be accounted for. This variation could cause an actually homogeneous treatment effect to appear to be heterogeneous and, thus, suggest a nonexistent moderator variable. (Alternatively, the effects of variation in treatment strength will be confounded with the real moderator variable.)

Consider an example. Suppose in a series of studies evaluating a new training method, the experimental group was supposed to get 10 hours of training in each study. Due to administrative and communications problems, however, the experimental people in some studies get 8, 9, 11, or 12 hours of training; although, within each study, each subject received exactly the same number of hours of training, only some of the studies hit exactly the desired 10 hours. If the mean across studies is 10 hours, then the mean effect size for the meta-analysis will not be affected. However, the variation in training time across studies will create additional variance in effect sizes beyond that created by sampling error. The formulas presented in this book do not correct for this. If the number of training hours is given in each study, this variable can be coded and analyzed as a moderator. However, this information would rarely be given because the deviations from 10 hours all represent errors in carrying out the training plan—errors that the experimenters themselves may not even be aware of.

In the example here, average treatment strength across studies was equal to the target value of 10 hours. This is what would be expected if the measurement error were random. If the mean were discrepant from the goal—say 9 hours instead of 10—then the mean effect size would be affected, as well as the variance. However, in this example, we assume a mean (expected value) of 0 for the measurement errors.
This form of measurement error is analogous to unintended differences between studies in range restriction in correlational studies, that is, differences in range restriction (or enhancement) that might appear despite the fact that researchers took special steps to obtain the same variation in all studies, just as the experimenters attempted here to have exactly the same treatment across studies. In many meta-analyses, the strength of treatment conditions will vary across studies, not because of measurement error, but because the different experimenters did not have a common goal for treatment strength to begin with. This condition is closely analogous to the naturally occurring range variation that occurs across correlational studies. (In the experimental studies, the control group is the same in all studies, anchoring the low end of the independent variables, but the high end of the independent variable varies from study to study, and hence the variance of the independent variable varies from study to study.) As noted and illustrated earlier, this problem can be addressed by a moderator analysis when the needed information on treatment strength is given in individual studies. However, this information will often not be given.

**Range Variation on the Dependent Variable**

The raw score treatment effect is determined by the nature of the treatment process. Thus, if the same process is used in different settings, it should stay about the same. However, the standard deviation of the study group is not determined by the treatment process but by the nature of the selection of the group in question. Thus, the study population might be more homogeneous in some settings than in others. The standardized treatment effect would vary correspondingly.

Consider an attitude change study done on a polarized political topic. Initial attitudes would be much more homogeneous in a group of Republicans than in a politically unselected population. Assume that the standard deviation among Republicans is only half the size of the standard deviation in a mixed population, say $\sigma = 50$ in the mixed population and $\sigma = 25$ for Republicans. If the change produced by the message is 10 points in raw score form, then a study done on a mixed population would produce a standardized effect size of $10/50 = .20$, while the same study done on a Republican population would produce a standardized effect size of $10/25 = .40$, a standardized effect size twice as large.

From the viewpoint of statistical power, there is a considerable advantage to doing a study using a more homogeneous population. Consider the political attitude example again. The investigator doing the study on a general population would have an effect size of $\delta = .20$, while the same study done on a Republican population would have an effect size of $\delta = .40$. Given a study sample size of $N = 100$, the statistical power for the study on a general population would be 17%, while the power on the homogeneous population would be 51%, three times higher.

The investigator studying the general population could have obtained a similar gain in power by breaking his data down into Republicans and Democrats and then properly merging the results from the two within-group comparisons. This is the
gain in power that results from analysis of covariance, or use of the “treatment by levels” design.

For purposes of meta-analysis, let us choose some population as a reference population. We want all effect sizes expressed in terms of that reference population. To do so, we must know the ratio of the standard deviation of the study population to the standard deviation of the reference population. Denote the standard deviations of the two populations by

\[
\sigma_P = \text{standard deviation of the reference population},
\]

\[
\sigma_S = \text{standard deviation of the study population}.
\]

The ratio of study to reference standard deviation is denoted \( u \), that is,

\[
u = \frac{\sigma_S}{\sigma_P}
\]

If the raw score treatment effect is the same in both populations, then the standardized treatment effect in the study population is given by

\[
\delta_S = \frac{\delta_P}{u}
\]

That is, the more homogeneous the study population in comparison to the reference population, the larger the study effect size.

To correct for range variation, we need merely use the preceding equation in reverse order, that is,

\[
\delta_P = u\delta_S
\]

In meta-analysis, this formula could be used to correct each of the study effect sizes to the same reference population value and, thus, eliminate differences in effect size due to differences in homogeneity. However, this correction requires that the same scale of measurement for the dependent variable be used in all studies. This is rarely the case, so this correction can usually not be made.

**Dichotomization of the Dependent Variable**

In some studies, a continuous dependent variable is dichotomized. For example, in research on the effect of a realistic job preview on subsequent turnover, most investigators do not use the natural dependent variable of tenure, the length of time the worker stays with the firm. Instead, they dichotomize tenure to create a binary “turnover” variable; for example, they might see if a worker stays more than 6 months or not. The loss of information inherent in dichotomization causes a reduction in the effect size and a corresponding loss in statistical power (MacCallum, Zhang, Preacher, & Rucker, 2002). Within a wide range of values, this artificial reduction in effect size can be corrected. However, within a single study, the statistical correction formula does not restore the higher level of statistical power.

Denote the treatment variable by \( T \) and denote the continuous dependent variable by \( Y \). Denote the dichotomized dependent variable by \( Y' \). The effect of the
dichotomization is to replace the correlation $r_{TY}$ by the correlation $r_{TY}'$, which is lower in magnitude. The statistical significance test is then done on the smaller $r_{TY}'$ with correspondingly lower power. What we seek is a correction formula that restores the value $r_{TY}$ to the value $r_{TY}'$. There is an approximate formula that works at the population level. Application of that formula at the sample level eliminates the systematic error in the correlation, but does not eliminate the larger sampling error that arises from the loss in information due to dichotomization. The formula works well in meta-analysis where the impact of sampling error is greatly reduced.

For a cross-sectional correlation, dichotomization of the dependent variable reduces the correlation by a product rule formula similar to that for attenuation due to error of measurement. The correction formula is known as that which creates a “bisetrical correlation” from a “point biserial correlation.” This formula does not work for treatment correlations because the dependent variable does not have a normal distribution. The treatment effect causes the distribution of the experimental group to be displaced from that of the control group. When the two groups are pooled, the combination distribution is not normal. To see this, consider the extreme case in which the treatment effect is 3 standard deviations in magnitude. The two distributions hardly overlap and the combined distribution is distinctly bimodal—one mode at each of the subgroup means.

However, we will show that the biserial correlation formula works quite well as an approximation over the usual range of effect sizes and distribution splits. This corresponds to the fact that the combined distribution is approximately normal unless the treatment effect is very large. Suppose that in the combined groups the proportion of people in the “high” split is $p$ while the proportion in the “low” split is $q = 1 - p$. For a normal distribution, there would be a $z$ value corresponding to such a split (although the combined distribution is not exactly normal). Call this value the “cutoff” value and denote it by $c$. The value of the normal density function or “normal ordinate” at $c$ is denoted $\phi(c)$. The attenuation in the treatment correlation is approximately given by the biserial attenuation formula

$$r_{TY}' = ar_{TY}$$

where

$$a = \phi(c)/\sqrt{pq}$$

The corresponding correction formula is the biserial formula

$$r_{TY} = r_{TY}' / a$$

The range over which the formula is accurate is shown in Table 6.3. Table 6.3 presents the comparison ratio for the actual continuous variable correlation and the attenuated dichotomous variable correlation corrected using the biserial correction formula. The ratio is in the order corrected/actual and is expressed as a percentage. For example, for a population continuous $d = .40$ and a median split on the combined population $p = .50$, the ratio is 101. That is, whereas the actual continuous treatment correlation is $r_{TY} = .20$, the corrected dichotomized correlation is $1.01(.20) = .202$, an error less than rounding error. The error is always less than rounding error for the range of values $-.51 < d < +.51$ and
Table 6.3  Comparison ratio of corrected/actual correlations—expressed as percentages—where the corrected correlation is the estimated correlation for the continuous dependent variable computed by correcting the dichotomous variable correlation using the biserial correction formula

<table>
<thead>
<tr>
<th>Combined proportion “high” on the dependent variable</th>
<th>.10</th>
<th>.20</th>
<th>.30</th>
<th>.40</th>
<th>.50</th>
<th>.60</th>
<th>.70</th>
<th>.80</th>
<th>.90</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>.20</td>
<td>100</td>
<td>100</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>.30</td>
<td>100</td>
<td>100</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>.40</td>
<td>99</td>
<td>100</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>101</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>.50</td>
<td>99</td>
<td>101</td>
<td>101</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>101</td>
<td>101</td>
<td>99</td>
</tr>
<tr>
<td>.60</td>
<td>98</td>
<td>101</td>
<td>102</td>
<td>103</td>
<td>103</td>
<td>103</td>
<td>102</td>
<td>101</td>
<td>98</td>
</tr>
<tr>
<td>.70</td>
<td>97</td>
<td>101</td>
<td>103</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>103</td>
<td>101</td>
<td>97</td>
</tr>
<tr>
<td>.80</td>
<td>96</td>
<td>101</td>
<td>103</td>
<td>105</td>
<td>105</td>
<td>105</td>
<td>104</td>
<td>101</td>
<td>96</td>
</tr>
<tr>
<td>.90</td>
<td>95</td>
<td>101</td>
<td>104</td>
<td>106</td>
<td>106</td>
<td>106</td>
<td>104</td>
<td>101</td>
<td>95</td>
</tr>
<tr>
<td>1.00</td>
<td>94</td>
<td>101</td>
<td>105</td>
<td>107</td>
<td>107</td>
<td>107</td>
<td>105</td>
<td>101</td>
<td>94</td>
</tr>
<tr>
<td>1.10</td>
<td>93</td>
<td>101</td>
<td>105</td>
<td>108</td>
<td>109</td>
<td>108</td>
<td>105</td>
<td>101</td>
<td>93</td>
</tr>
</tbody>
</table>

NOTE: The statistic $d$ is the population effect size for the continuous variable, that is, approximately twice the value of the population continuous variable treatment correlation.

$.09 < p < .91$, the range of values in most current meta-analyses. For the most extreme case in Table 6.3, $d = 1.10$ and $p = .90$, the actual correlation is $.48$ and the corrected correlation is $.93(.48) = .45$, an error that is visible but still not large in practical terms.

**Imperfect Construct Validity in the Dependent Variable**

Suppose there is some systematic error in the measurement of the dependent variable; that is, we measure a dependent variable that is different to some extent from the intended dependent variable. What effect will this have on the effect size and can it be corrected? A full treatment of this problem requires considerable knowledge of path analysis and knowledge of the nature of the dependent variable and its relationships with other variables. However, there are certain common cases that are relatively straightforward.

The most common case is the use of a dependent variable that is an indirect measure of the desired dependent variable. For example, a good assessment of a juvenile delinquency treatment program would require an objective assessment of the subsequent behavior of the clients. Instead, investigators must often rely on indirect measures such as the subsequent arrest record. Figure 6.2 shows the assumed path model of the relationships between the two measures of behavior and the treatment variable.

Let $Y$ be the measure of the client’s actual delinquent behavior and let $Y'$ be the arrest record. The desired treatment effect correlation $r_{TY}$ is related to the observed treatment correlation by the product rule

$$r_{TY'} = r_{TY} r_{YY'}$$

(6.12)
Figure 6.2 Path model for the relationship among the delinquency treatment program, the desired measure of actual posttreatment behavior, and the observed posttreatment arrest record

\[ r_{TY} \rightarrow Y \rightarrow Y' \]

Legend:
- T = Delinquency Treatment Variable
- \( Y' \) = Posttreatment Behavior
- Y = Posttreatment Arrest Record

If the correlation between behavior and arrest were only \( r_{YY'} = .30 \), then the treatment correlation would be attenuated to

\[ r_{TY'} = .30r_{TY}, \]

that is, attenuated by 70%. In this case, the observed correlation could be corrected by reversing the algebraic equation

\[ r_{TY} = r_{TY'} / r_{YY'} \quad (6.13) \]

The corrected \( d \) statistic would then be obtained by transforming this corrected correlation. In the delinquency example, the correction would be

\[ r_{TY} = r_{TY'} / .30 \]

The observed correlation must be more than tripled in this case to correct for the imperfect construct validity. If the observed \( d \) statistic were \( d_{Y'} = .12 \), then \( r_{TY'} = .06 \), which corrects to \( r_{TY} = .06 / .30 = .20 \) and, hence, to \( d_Y = .40 \).

Although the treatment correlation, or \( d \) statistic, can be corrected to eliminate the systematic reduction in the correlation produced by imperfect construct validity, the effect of increased sampling error cannot be corrected. The confidence interval around the corrected correlation is \( 1 / .30 = 3.33 \) times as wide as that for the uncorrected correlation, reflecting the increased sampling error caused by the correlation. The significance test for the corrected effect size is algebraically equivalent to the significance test for the uncorrected effect size and, hence, has the same \( p \) value.

Imperfect construct validity does not always reduce the size of the effect size. Consider social skills training for supervisors. Assessment of the training program would ideally require measuring the interactive skills of the trainee after the program. Instead, the only available measure might be a measure of how well the person mastered the training program. However, mastery of the material is only antecedent to behavior change; it may take time or special experience for the trainee to put that learning into operation. The path model for this hypothesis is shown in Figure 6.3.

The desired treatment correlation \( r_{TY} \) and the observed treatment correlation \( r_{TY'} \) are related by the product rule

\[ r_{TY} = r_{TY'} r_{Y'Y}. \quad (6.14) \]
Figure 6.3  Path model for the assumed relationship among the social skills training of supervisors, program mastery, and subsequent interpersonal behavior on the job

Legend:

\( T \) = Training Treatment Variable
\( Y' \) = Training Learning Measure
\( Y \) = Posttreatment Social Behavior

If the correlation between the cognitive learning measure and subsequent social behavior was only \( r_{Y'Y} = .30 \), then the desired correlation \( r_{TY} \) would be lower than the observed correlation \( r_{TY'} \):

\[ r_{TY} = .30 r_{TY'} \]

This product rule is itself the correction formula for the treatment correlation. The correction for the \( d \) statistic is obtained by transforming the corrected treatment correlation to a \( d \) value.

Imperfect Construct Validity in the Treatment Variable

Imperfect construct validity in the treatment variable is a confounding of the intended treatment effect with an effect due to some other causal agent. This problem can be attacked using path analysis (Hunter, 1986, 1987), and, under certain conditions, it can be corrected in a manner that could be used in meta-analysis. Correction of confounding requires the use of a multiple dependent variable design where an intervening variable is observed that measures the process induced by the confounding causal agent. The desired correlation is then the partial correlation between the treatment variable and the dependent variable with the intervening variable held constant. This approach is a rigorous, quantitative replacement for doing an analysis of covariance with the intervening variable used as a concomitant variable (Hunter, 1988). Detailed treatment of this method is beyond the scope of the present book.

Bias in the Effect Size (\( d \) Statistic)

The effect size statistic is subject to a statistical phenomenon known by the forbidding title "bias." For sample sizes greater than 20, bias is trivial in magnitude. However, there have been papers arguing that a major problem with meta-analysis is the use of "biased methods." This section will present a correction for bias, although the primary intent of the section is to show that the bias is of trivial magnitude.
Consider a set of perfectly replicated studies, all with the same sample size and the same population effect size. The average sample effect size will differ slightly from the population effect size. This discrepancy is called “bias” in the statistical estimation literature. The size of the bias depends on the sample size. The bias is in different directions for the effect size statistic $d$ and the treatment correlation $r$. The average $d$ is slightly larger than $\delta$ while the average correlation is slightly smaller than $\rho$.

A complete treatment of bias in the treatment correlation is given by Hunter, Schmidt, and Coggin (1996), who showed that the bias is trivial in all but very small sample studies and small for small-sample studies. Some authors have suggested using the Fisher’s $z$ transformation to reduce the bias, but the bias using Fisher’s $z$ turns out to be even larger, although in the opposite direction. The bias in the treatment correlation is given by

$$E(r) = a\rho$$

where

$$a = 1 - (1 - \rho^2)/(2N - 2)$$

How small is this bias? Consider perfectly replicated studies with a population correlation of $\rho = .20$ and a sample size of $N = 100$. The multiplier for the average observed correlation would be $a = 1 - .00485 = .995$. The average correlation would be .199 instead of .200. The trivial size of the bias in the correlation is the reason that bias has traditionally been ignored. However, for very small sample size meta-analyses (average sample size less than 10), or for very fastidious analysts, it is possible to correct the observed treatment correlation for bias. The nonlinear correction corresponding to the preceding equation would be

$$r' = r/a$$

where

$$a = 1 - (1 - r^2)/(2N - 2)$$

If the meta-analysis is studying treatment correlations below .70 in magnitude, then the correction is very closely approximated by the linear correction

$$r' = r/a$$

where

$$a = (2N - 3)/(2N - 2)$$

The linear correction has the advantage that it can be applied after the meta-analysis. Just divide both the estimated mean and the estimated standard deviation of population correlations by the multiplier computed with $N$ set at the average sample size. Note that the corrected correlation will be trivially larger than the uncorrected correlation.
A complete treatment of bias in the \( d \) statistic is given in Hedges and Olkin (1985). Although the bias is trivial for all but very small sample studies, they recommend routine correction for that bias. Indeed, they use the symbol \( d \) only for the corrected statistic. The bias in \( d \) is approximately given by

\[
E(d) = a \delta
\]

where

\[
a = 1 + \frac{3}{(4N - 12)}
\]

(6.21)

How large is this bias? Consider a study with a textbook-sized effect \( d = .40 \) based on a sample size of \( N = 100 \). The multiplier is \( a = 1 + .0077 = 1.0077 \). The average study effect size would be .403 instead of .400. On the other hand, consider a study done with an extremely small sample size, say, \( N = 10 \) (five subjects in each group). The multiplier would be \( a = 1 + .107 = 1.107 \) and the average effect size would be .443 rather than .400. The correction is straightforward; just divide by \( a \):

\[
d' = d/a
\]

(6.22)

Note that this is a linear correction. Thus, it could be applied after the meta-analysis. Just divide both the estimated mean and the estimated standard deviation of corrected \( d \) values by the multiplier \( a \). Again, the \( N \) used should be the average \( N \). The corrected effect sizes will be slightly smaller than the uncorrected effect sizes.

**Recording, Computational, and Transcriptional Errors**

Meta-analysis will inevitably include some studies with bad data. There might be a recording error in gathering the primary data or in entering it into the computer. The study effect size could be erroneous because of computational error or an error in algebraic sign in the effect size. Finally, error can arise in transcription: from computer output to analyst table, from analyst table to manuscript table, from manuscript table to published table. Some have even suggested that a meta-analyst might miscopy a figure, but it is well-known that meta-analysts do not make errors.

Study results should be examined for extreme outliers. This can eliminate the most extreme cases of bad data. However, keep in mind the caveats about outlier analysis in meta-analysis presented in Chapters 3 and 5. Also, even with an outlier analysis, smaller erroneous effect sizes will not be detectable. Thus, any meta-analysis with a very large number of effect sizes will usually have at least a few bad data points.

Because bad data cannot be completely avoided, it is important to consider variation in study results with some caution. A certain percentage of the original observed variation will be due to bad data. A larger proportion of the residual variation—that left after all other study artifacts have been corrected—may be due to bad data.
Multiple Artifacts and Corrections

Unfortunately, there is no rule that says that study results can be distorted by only one artifact. Sampling error will be present in all studies. Error of measurement in the dependent variable will be present in all studies. Imperfect control of the nominal treatment variable will be unavoidable in most studies. Thus, a meta-analysis free of artifact will usually require that effect sizes be corrected for a number of sources of error.

Other than the removal of the most extreme outliers, there is no correction for bad data. Sampling error acts differently from other artifacts in that it is (1) additive and (2) unsystematic. Although a confidence interval for the effect size provides an unbiased estimate of potential sampling error for the single study, there is no correction for sampling error at the level of the single study. We will consider correcting the problem of sampling error in meta-analysis in the next chapter.

The artifacts other than sampling error and bad data are systematic in nature and, thus, potentially correctable. The key is to have the necessary information about the size of the artifact process (e.g., knowledge of the extent of unreliability or the extent of range restriction or the extent of imperfect construct validity). You can correct each artifact where there is adequate artifact information, be this one artifact or seven. Every artifact left uncorrected results in a corresponding underestimate of the true effect size. Furthermore, variation in uncorrected artifacts across studies looks like true variance in treatment effect. This may create the appearance of across-setting variation where there is none. If there is true variation in treatment effect, then uncorrected artifacts mask the true differences. That is, variation in apparent effect size due to uncorrected artifacts may override the differences due to true moderator variables and, thus, make it difficult to identify the true moderator variable by examining the studies after the fact.

Correction for artifacts is not difficult if the information on each artifact is given. Consider an example: a study of social skills training for first-line supervisors conducted in a factory setting. The measure of job performance is performance ratings by the immediate manager of each supervisor on a single 100-point graphic rating scale. Because the investigator doubted that performance ratings are measured on a ratio scale (a fact), the investigator decided that parametric statistics could not be used (a statistical error on his part). So the investigator decided to do a sign test on the data. The combined group data were split at the median and a chi-square test was run comparing the proportion of above-average performance among the supervisors with and without the skills training. Assume the treatment effect for training on consensus performance ratings is \( d_{TY} = .40 \). The true population treatment correlation is thus \( r_{TY} = .20 \). The inter-rater reliability of performance ratings by a single supervisor on a single rating scale averages .28 (Hunter & Hirsh, 1987; King, Hunter, & Schmidt, 1980). Thus, the effect size is reduced from \( d_{TY} = .40 \) to

\[
d_{TY} = \sqrt{.28} d_{TY} = .53 d_{TY} = .53 (.40) = .21
\]

For a study with a total sample size of 100, this would reduce statistical power from an already low 51% to a very low 18%. The effect of a dichotomization using
a median split for a $d$ value less than .50 is simply to decrease the value by 20%, that is,

$$d_{TY'} = .80d_{TY} = .80(.21) = .17$$

This reduces the statistical power from a very low 18% to an even lower 13%. That is, the two artifacts together reduce the effect size from .40 to .17, less than half its proper value. The statistical power is reduced from an already undesirable 51% to only 13%, an increase in the error rate for the significance test from 49% to 87%.

The systematic effect of the artifacts can be eliminated from the study by correction formulas. We have

$$d_{TY} = d_{TY'}/.53 = (d_{TY'}/.80)/.53 = d_{TY''}/.424$$

That is, if the artifacts reduce the size of the effect by 58%, we can restore the value by dividing by the corresponding factor .42. However, while this correction eliminates the systematic error in the effect size, it does not eliminate the increased sampling error. A proper statistical test on the corrected effect size gives the same $p$ values as the test on the uncorrected effect size. Thus, statistical correction formulas do not restore the lost statistical power.

Consider the preceding example in abstract form. The effect of the first artifact is to multiply the effect size by a multiplicative factor $a_1$ as in

$$d' = a_1d$$

The effect of the second artifact is to multiply by a second factor $a_2$ as in

$$d'' = a_2d'$$

The impact of the two factors is

$$d'' = a_2d' = a_2(a_1)d = a_1a_2d$$

namely, to multiply the effect size by a multiplicative factor $a$, which is the product of the two separate artifact multipliers. That is,

$$d'' = ad$$

where

$$a = a_1a_2$$

The effect size is then restored to its original value by the correction formula

$$d = d''/a \quad (6.23)$$

The preceding example is typical of correction for multiple artifacts in effect sizes of moderate size. Each artifact reduces the effect size by a multiplicative factor. The net effect is to reduce the effect size by a multiplicative factor that is the product of the separate multipliers. That is, the attenuating effect of several
artifacts is to reduce the effect size by the product of the separate attenuating factors. The corresponding correction restores the effect size to its original size by dividing by the attenuation multiplier.

There is one caveat to the preceding discussion: It is exactly true for the treatment correlation but only an approximation for the effect size $d$. If a treatment effect is so large that the approximation $d = 2r$ breaks down, then the multiplicative formula should be applied only to the treatment correlation. The treatment effect $d$ can then be computed by the usual conversion formula.

For example, consider the social skills training example for a very large treatment effect, say, $d = 1.50$. The observed treatment correlation satisfies the attenuation equation

$$ r_{TY}'' = .42 r_{TY} $$

The effect size $d = 1.50$ corresponds to a treatment correlation of $.60$. The artifacts reduce this correlation from $.60$ to

$$ r_{TY}'' = .42(.60) = .25 $$

This attenuated treatment correlation corresponds to an attenuated effect size of $d = .52$. The attenuation in the $d$ statistic is from $1.50$ to $.52$, which is by a factor of $.35$ rather than the factor $.42$ for the treatment correlation.

The treatment of multiple artifacts is very straightforward for the treatment correlation. The reduction in the size of the treatment correlation by a series of artifacts is effected by multiplication of the correlation by a total multiplier that is the product of the separate artifact attenuation multipliers. The treatment correlation can be restored to its proper size by correcting for those artifacts by dividing by the attenuation multiplier.

For the $d$ statistic, correction for artifacts is only slightly more difficult. If the unattenuated population $d$ value falls in the moderate range $-.50 < d < +.50$, then the formulas for the treatment correlation apply directly to the $d$ statistic. If the population effect size falls outside the moderate range, then the $d$ value can be converted to a correlation, the correlation can be corrected, and the corrected correlation can be converted to a corrected $d$ value.