CHAPTER 7

The Importance of Reliability

Throughout this book, we have emphasized the fact that psychological measurement is crucial for research in behavioral science and for the application of behavioral science. As a cornerstone of a test’s psychometric quality, reliability is a fundamental issue in understanding and evaluating the quality of psychological measurement. The previous two chapters detailed the conceptual basis of reliability and the procedures used to estimate a test’s reliability. In this chapter, we articulate the important roles that reliability plays in the applied practice of behavioral science, in behavioral research, and in test construction and refinement.

Applied Behavioral Practice: Evaluation of an Individual’s Test Score

Psychological test scores are often used by psychologists and others to make decisions that have important effects on people’s lives. For example, as mentioned in the first chapter of this book, intelligence test scores can be used by courts to determine eligibility for the death sentence for convicted murderers. This may be an extreme example of how test scores can affect our lives, but it illustrates the importance of having reliable scores. It would be tragic, to say the least, if someone were sentenced to death based on an unreliable intelligence test score.

There are uncounted other, albeit less dramatic, instances in which the reliability of scores on psychological tests can have an impact on the lives of ordinary people. Children are often removed from standard academic classrooms and assigned to special classes based on intelligence and achievement test scores. Similarly, tests such as the SAT and Graduate Record Examination (GRE) are used to make decisions about college admissions, and employers often use tests to make hiring and promotion decisions. Classroom instructors may not give the problem of test reliability much thought when they give their class examinations, but scores on those examinations can have an influence on students’ futures.
A test's reliability has crucial implications for the quality of decisions that are made on the basis of an individual's test scores. Recall that we can never know an individual's "true" level on an unobservable psychological construct. For example, we can never know a person's true level of intelligence or capacity for college achievement. Thus, we use psychological test scores to indicate or estimate an individual's true level of some psychological attribute.

Because test scores are only estimates of people's actual psychological characteristics and because decisions about persons' lives are often based partly on these scores, we must evaluate the precision of the score obtained by any particular individual on a test. That is, we would like to be able to gauge the precision or accuracy of an individual's test score as an estimate of the individual's psychological attribute. As we will see, the reliability of test scores can be used to calculate information that will help us evaluate the precision of particular test scores.

Two important sources of information can help us evaluate an individual's test score. First, a point estimate is a specific value that is interpreted as a "best estimate" of an individual's standing on a particular psychological attribute. As we will discuss, there are two ways of obtaining a point estimate for an individual.

The second source of information that helps us evaluate an individual's test score is a confidence interval. A confidence interval reflects a range of values that is often interpreted as a range in which the true score is likely to fall. The logic of a confidence interval is based on the understanding that an observed score is simply an estimate of a true score. Because of measurement error, the observed score may not be exactly equal to the true score.

The confidence interval around a particular score gives us an idea of the score's accuracy or precision as an estimate of a true score. If we find that an individual's score is associated with a narrow confidence interval, then we know that the score is a fairly precise point estimate of the individual's true score. However, if we find instead that an individual's score is associated with a wide confidence interval, then we know that the score is an imprecise or inaccurate point estimate of the individual's true score. We will see that these values—point estimates and confidence intervals—are directly affected by test score reliability.

**Point Estimates of True Scores**

Two kinds of point estimates can be derived from an individual's observed test score, representing the best single estimate of the individual's true score. The most widely used point estimate is based solely on an individual's observed test score. When an individual takes a test at a given point in time, his or her observed score can itself be used as a point estimate. For example, if you give someone a self-esteem test, his or her score on the test can be seen as a point estimate of his or her true self-esteem score.

The second type of point estimate, sometimes called an adjusted true score estimate, takes measurement error into account. Once again, recall that an individual's observed score on any given test is affected by measurement error. Because testing is never perfect, an individual's test score may be somewhat inflated or deflated by
momentary factors, such as fatigue, distraction, and so on. Therefore, an individual's test score at one time is artificially high or low compared with the score that the individual would likely obtain if he or she took the test a second time. As a matter of fact, if an individual took the same test on two occasions, then he or she would likely obtain two observed scores that are at least slightly different from each other. Both of those observed test scores could be considered point estimates of the individual's true score. With an understanding of reliability and the nature of measurement error, we can use an individual's observed score from one testing occasion to estimate the results that we might obtain if we tested the individual repeatedly. This produces an adjusted true score estimate, which reflects an effect called regression to the mean.

Regression to the mean refers to the likelihood that on a second testing, an individual's score is likely to be closer to the group mean than was his or her first score. That is, if an individual's observed score is above the mean on the first testing occasion, then he or she is likely to score somewhat lower (i.e., closer to the mean) on the second testing occasion. Similarly, if an individual's observed score is below the mean on the first testing occasion, then he or she is likely to score somewhat higher (i.e., closer to the mean) on the second testing occasion. This prediction is again based on the logic of classical test theory (CTT) and random measurement error. In Chapter 5, we learned that measurement error is random and likely to affect all test scores to some degree—artificially inflating some scores (that end up relatively high) and artificially deflating some scores (that end up relatively low).

The adjusted true score estimate is intended to reflect the discrepancy in an individual's observed scores that is likely to arise across repeated testing occasions. The size and direction of this discrepancy is a function of three factors: (1) the reliability of the test scores, (2) the size of the difference between the individual's original observed test score and the mean of the test scores, and (3) the direction of the difference between the original score and the mean of the test scores. These factors can be used to calculate the adjusted true score estimate through the following equation:

\[ X_{est} = \bar{X}_o + R_{xx}(X_o - \bar{X}_o), \]

where \( X_{est} \) is the adjusted true score estimate, \( \bar{X}_o \) is the test's mean observed score, \( R_{xx} \) is the reliability of the test, and \( X_o \) is the individual's observed score. For example, imagine that you have scores from a multiple-choice exam given to a class. There are 40 questions on the exam, and the exam mean is 30. Assume that the exam has an estimated reliability of .90 (this would be a very high reliability for most class examinations). If a student had a score of 38 on the exam, then his or her estimated true score would be

\[ X_{est} = 30 + .90(38 - 30), \]

\[ = 37.2. \]

Notice that the estimated true score (37.2) is closer to the mean (30) than was the initial observed score (38). Thus, the adjusted true score attempts to account for the likely occurrence of regression to the mean.
There are at least two important points to note about the adjusted true score estimate, in relation to the observed score. First, test reliability influences the difference between the estimated true score and the observed score. Specifically, as reliability decreases, the difference between the adjusted true score estimate and the observed score increases. That is, poorer reliability produces bigger discrepancies between the estimated true score and the observed score. This reflects the fact that regression to the mean is more likely to occur (or is likely to be more substantial) when a test's scores are affected heavily by measurement error. For example, assume that the class test's reliability is only .50 and we computed the adjusted true score estimate for an individual with an observed score of 38:

\[ X_{est} = 30 + .50(38 - 30), \]

\[ = 34. \]

Thus, for an individual with a test score of 38, the predicted effect of regression to the mean is 4 points (38 − 34 = 4) for a test with poor reliability but less than 1 point (38 − 37.2 = .8) for a test with strong reliability.

A second important implication of the adjusted true score estimate is that the observed score's extremity influences the difference between the estimated true score and the observed score. Specifically, the difference will be larger for relatively extreme observed scores (high or low) than for relatively moderate scores. For example, let us compute the adjusted true score estimate for an individual with an observed score of 22 (i.e., an observed score that is 8 points below the mean of 30) on a test with a reliability of .90:

\[ X_{est} = 30 + .90(22 - 30), \]

\[ = 22.8. \]

Note that the adjusted true score estimate is 0.8 points closer to the mean than the observed score in this case. Now, let us compute the adjusted true score estimate for an individual with an observed score of 27 (i.e., a less extreme observed score that is only 3 points below the mean of 30):

\[ X_{est} = 30 + .90(27 - 30), \]

\[ = 27.3. \]

Note that this adjusted true score estimate is only 0.3 points closer to the mean than the observed score. Thus, the adjustment was more substantial for the relatively extreme observed score (i.e., 22) than it was for the less extreme observed score (i.e., 27).

Although the ideas of an adjusted true score estimate and regression to the mean are pervasive features of most attempts to evaluate individual scores on a test (e.g., see Wechsler, 2003a, 2003b), there are reasons to approach these ideas with caution. First, except in the case where we intend to predict a person's score on a subsequent test or form a confidence interval (see below), there seems to be little reason to correct observed scores by adjusting them for regression to the mean. Indeed, Nunnally and Bernstein (1994) state that “one rarely estimates true scores [in the
“adjusted true score” sense that we discuss] in the applied assessment of static constructs” and that “it is easier to interpret the individual’s obtained score” (p. 260). Second, although most psychologists seem to think that regression to the mean is, in the long run, a mathematical certainty, Rogosa (1995) has shown that there are circumstances in which it will not occur. Nevertheless, as we will see when we discuss true score confidence intervals, it is common practice to convert observed scores to adjusted true score estimates.

Confidence Intervals

In applied testing situations, point estimates of an individual’s score are usually reported along with confidence intervals. Roughly speaking, confidence intervals reflect the accuracy or precision of the point estimate as reflective of an individual’s score. For example, we might administer the Wechsler Intelligence Scale for Children (WISC) to a child and find that the child obtains a score of 106. Taking this observed score as an estimate of the child’s true score, we might calculate a confidence interval and conclude that we are “95% confident that the individual’s true IQ score falls in the range of 100–112” (Wechsler, 2003b, p. 37). The width of a confidence interval (e.g., a 12-point range) reflects the precision of the point estimate. You will probably not be surprised to learn that this precision is closely related to reliability—tests with high reliability provide estimates that are relatively precise.

The link between reliability and the precision of confidence intervals is made through the standard error of measurement (\(se_m\)). As discussed in Chapter 5, the \(se_m\) represents the average size of the error scores that affect observed scores. The larger the \(se_m\), the greater the average difference between observed scores and true scores. Thus, the \(se_m\) can be seen as an index of measurement error, and it is closely linked to reliability. In fact, Equation 5.16 presented the exact link between the standard error of measurement (\(se_m\)), reliability (\(R_{xx}\)), and the standard deviation of a test’s observed scores (\(s_o\)):

\[
se_m = s_o \sqrt{1-R_{xx}}.
\]

Once we have estimated the standard error of measurement for a set of test scores, we can compute a confidence interval around an individual’s estimated true score. To report a 95% confidence interval around that score, we would use the following equation:

\[
95\% \text{ confidence interval} = X_{ort} \pm (1.96)(se_m),
\]

where \(X_{ort}\) is the adjusted true score estimate (i.e., a point estimate of the individual’s true score) and \(se_m\) is the standard error of measurement of the test scores. The final component of this equation (1.96) reflects the fact that we are interested in a 95% confidence interval rather than a 90% interval or any other “degree of confidence” (we will address alternate “degrees of confidence” later). Some readers—particularly those who have a background in statistical significance testing—might recognize this value as being associated with a probability of .95 from the standard normal distribution.
For example, imagine an individual's observed score is 38, based on a test with a mean observed score of $X_o = 30$, a standard deviation (of observed scores) of $s_o = 6$, and an estimated reliability of $R_{xx} = .90$. From calculations earlier, we know that her adjusted true score estimate is $X_{est} = 37.2$. From Equation 5.16, we estimate the standard error of measurement as

$$se_m = 6\sqrt{1-.90},$$

$$= 1.90.$$

Based on Equation 7.2, for our test, the 95% confidence interval is 28.3 to 35.7:

$$\text{95% confidence interval} = 37.2 \pm (1.96)(1.90),$$

$$= 37.2 \pm 3.7,$$

$$= 33.5 \text{ to } 40.9.$$ Using the logic expressed by the above quote from Wechsler, we might interpret this result as indicating that we are 95% confident that the individual's true score falls in the range of 33.5 to 40.9.

As mentioned earlier, the precision of a true score estimate is closely related to reliability. Briefly put, highly reliable tests produce narrower confidence intervals than less reliable tests. We just saw that for our highly reliable test ($R_{xx} = .90$), the $se_m$ was 1.90 and the confidence interval had a range of 7.4 points ($40.9 - 33.5 = 7.4$). The size of this range reflects the precision of the confidence interval—the smaller or narrower the interval, the more precise the observed score is as an estimate of the true score. Although highly reliable tests produce narrow intervals, less reliable tests will produce wider (i.e., larger) confidence intervals, reflecting a less precise estimate of the true score. For example, let us imagine that our test had the same observed score standard deviation as our previous example ($s_o = 6$) but a lower reliability (say only $R_{xx} = .50$). In this case of a test with poor reliability, the standard error of measurement would be 4.2:

$$se_m = 6\sqrt{1-.50},$$

$$= 4.24.$$

Note that this $se_m$ is larger than it was for the previous example, in which reliability was .90 and the $se_m$ was only 1.90. As we have seen, the $se_m$ has a direct effect on the confidence interval. So in the case of our low-reliability test, the 95% confidence interval around an adjusted true score estimate of 37.2 is a relatively wide range of 28.9 to 45.5:

$$\text{95% confidence interval} = 37.2 \pm (1.96)(4.24),$$

$$= 37.2 \pm 8.3,$$

$$= 28.9 \text{ to } 45.5.$$ Thus, the test with poor reliability produced a much less precise (i.e., wider) confidence interval than the test with high reliability. Specifically, the test with $R_{xx} = .50$ produced an interval of 16.6 points ($45.5 - 28.9 = 16.6$), but we saw that the test with $R_{xx} = .90$ produced an interval of only 7.4 points. It is a much stronger
and more precise statement to say that “we are 95% confident that an individual’s true score lies between 33.5 and 40.9” than it is to say that “we are 95% confident that the individual’s true score lies anywhere all the way from 28.9 to 45.5.”

**Debate and Alternatives**

This section has outlined one way that has been recommended for reporting individual scores and confidence intervals around those scores. However, it is certainly not the only way that you might encounter. Perhaps more commonly, you will see observed scores reported, and you will see confidence intervals (if they are reported at all) computed around those observed scores, and you will see the intervals interpreted as above.

There is considerable debate and variations in the ways in which confidence intervals are computed and integrated with true score estimates. Confidence intervals can be computed for various degrees of confidence (e.g., 99% or 90% or 68% instead of 95%), they can be computed by using either the standard error of measurement or a related value called the standard error of estimate (which is also affected by reliability), and they can be applied to either observed score estimates of true scores or adjusted true score estimates (as described in the previous section).

These various alternatives, in turn, have implications for the exact interpretation of the confidence intervals. According to true score theory, observed scores are distributed normally around true scores. Because an observed score is the best estimate of a true score, the observed score represents the mean of this distribution. In our example, an adjusted true score estimate of 37.2 may lie within a 95% confidence interval that ranges from 33.5 to 40.9, but what does it mean to say that the score is in this confidence interval? Perhaps the most widely offered answer to this question is, as we illustrate, that “there is a 95% chance that the true score falls within the confidence interval.” Another way to say the same thing is “The probability is .95 that the confidence interval contains the true score.” These statements might be interpreted in two different ways. They might mean that there is a 95% chance that a person’s true score will fall in the interval on repeated testing with the same or parallel tests, or it might mean that if you had many people with the same true score take the same test, 95% of their observed scores would fall in the interval. However, disagreement exists over such interpretations.

For example, referring to the typical computation of confidence intervals, Dudek (1979) objects to interpretations such as “There is a 95% chance that the true score falls within the confidence interval” because answers of this type imply that true scores are deviating around an observed score. He suggests such interpretations would require use of the adjusted true score estimate, along with a different version of the standard error. We have sympathy for this view, but in most cases, when confidence intervals are computed as illustrated above, they are interpreted (again for better or for worse) in a way that suggests that true scores are falling somewhere in the confidence interval.

Although such variations emerge in some applications of psychological testing, details of these variations are well beyond the scope of our current discussion. Interested readers are encouraged to refer to other sources for more details, including...

**Summary**

For our purposes, the most important general message from this section is that reliability affects the confidence, accuracy, or precision with which an individual's true score is estimated. That is, reliability affects the standard error of measurement, which affects the width of a confidence interval around an individual's estimated true score. Poor reliability produces scores that are imprecise reflections of individuals' true psychological traits, skills, abilities, attitudes, and so on. Good reliability produces scores that are much more precise reflections of individuals' true psychological attributes. If we truly want to use psychological tests to make decisions about individuals, then those tests should have strong reliability.

Again, the issues associated with estimated true scores and true score intervals might seem abstract and esoteric, but they can have important consequences in applied settings in which test scores are used to make decisions about the lives of individual people. For example, children are often classified as having mental retardation if they have an intelligence test score below 70. We know, however, that any IQ score will have some degree of unreliability associated with it (although the reliability of scores on standard, individually administered intelligence tests is very high). The degree of test score unreliability should influence your interpretation of an observed score; to what extent does an observed score reflect a child's true score? Imagine that a child has a tested IQ score of 69. How confident would you be that the child's true score is below 70, and how likely is it that, on a second testing, the child's tested score might be greater than 70? We know that, in all likelihood, if this child is given a second intelligence test, the child's IQ score will increase because of regression to the mean. At what point do we take these factors into consideration, and how do we do so when making a decision about the child's intellectual status?

It is imperative that those making these types of decisions recognize the problems associated with the interpretation of psychological test scores. Our hope is that you recognize the problem and appreciate the fact that reliability has a fundamental role in it.

**Behavioral Research**

Reliability has important implications for interpreting and conducting research in the behavioral sciences. The interpretability of research in areas such as psychology and education hinges on the quality of the measurement procedures used in the research. In this section, we explain how reliability and measurement error affect the results of behavioral research. Awareness of these effects is crucial for interpreting behavioral research accurately and for conducting behavioral research in a productive way.
Reliability, True Associations, and Observed Associations

Earlier in this book, we discussed the importance of understanding associations between psychological variables (see Chapter 3). That is, one of the most fundamental goals of research is to discover the ways in which important variables are related to each other. For example, researchers might want to know whether SAT scores are associated with academic performance or whether personality similarity is associated with relationship satisfaction or whether “dosage of medication” is associated with decreases in depressive affect. Thus, knowing the direction and magnitude of the associations between variables is a central part of scientific research.

Psychological scientists usually rely on several basic ways of quantifying the association between variables. In terms of psychometrics, the most common way of doing this is through a correlation coefficient (again, see Chapter 3). Thus, in the following discussion, we focus mainly on the correlation coefficient as a way of explaining the importance that reliability has on behavioral research. However, it is important to realize that researchers often use other statistics to reflect the association between variables. For example, experimental psychologists are more likely to use statistics such as Cohen’s $d$ or $\eta^2$ than they are a correlation coefficient. We will touch on these statistics briefly here.

According to CTT, the correlation between observed scores on two measures (i.e., $r_{XY_{oo}}$) is determined by two factors: (1) the correlation between the true scores of the two psychological constructs being assessed by the measures (i.e., $r_{XY_{tt}}$) and (2) the reliabilities of the two measures (i.e., $R_{XX}$ and $R_{YY}$). Specifically,

$$r_{XY_{oo}} = r_{XY_{tt}} \sqrt{R_{XX} R_{YY}}.$$

Equation 7.3 is the key element of this section, with many important implications for research and applied measurement. Before we discuss those implications, we will explain how Equation 7.3 follows logically from CTT (including the assumption that error scores are random and therefore uncorrelated with true scores and other sets of error scores, but cf. Charles, 2005, and Nimon, Zientek, & Henson, 2012).

Recall again from Chapter 3 (Equation 3.5) that the correlation between two variables ($r_{XY}$) is the covariance divided by two standard deviations:

$$r_{XY} = \frac{c_{XY}}{s_X s_Y}.$$

In terms of observed scores, the correlation between scores from two measures is

$$r_{XY_{oo}} = \frac{c_{XY_{oo}}}{s_X s_Y}.$$

We will think about the numerator of this equation for a moment. Recall from Chapter 5 that, according to CTT, observed scores are composite variables (i.e., $X_o = X_t + X_e$ and $Y_o = Y_t + Y_e$). Therefore, the covariance between two sets of observed
scores (i.e., observed scores on X and observed scores on Y) can be seen as the covariance between two composite variables. Following the example outlined in Chapter 3’s discussion (e.g., Equation 3.7) of the covariance between composite variables, the covariance between X and Y (i.e., $c_{X,Y}$) is

$$c_{X,Y} = c_{x,t} + c_{x,e} + c_{x,t} + c_{x,e},$$

where $c_{x,t}$ is the covariance between true scores on test X and true scores on test Y, $c_{x,e}$ is the covariance between true scores on test X and error scores on test Y, $c_{x,t}$ is the covariance between error scores on test X and error scores on test Y, and $c_{x,e}$ is the covariance between error scores on test X and error scores on test Y. By definition, error scores occur as if they are random. Therefore, error scores are uncorrelated with true scores, and error scores on test X are uncorrelated with error scores on test Y. Consequently, the three covariances that include error scores are equal to 0, which means that the covariance between observed scores reduces to the covariance between true scores ($c_{x,t} = c_{x,t}$). Thus, returning to Equation 7.4, the correlation between two sets of observed scores is

$$r_{x,y} = \frac{c_{x,y}}{s_x s_y}.$$  (7.5)

Next, we will think about the denominator of this equation. Recall from Chapter 5 that variability in a test’s observed scores (e.g., $s_x$ and $s_y$) is related to the test’s reliability. Specifically, reliability can be defined as the ratio of true score variance to observed score variance:

$$R_{xx} = \frac{s_x^2}{s_{xt}} \quad \text{and} \quad R_{yy} = \frac{s_y^2}{s_{yt}}.$$  

Rearranging these, we can express the observed standard deviations as a function of reliability and standard deviations of true scores:

$$s_x = \frac{s_{xt}}{\sqrt{R_{xx}}}$$  (7.6a) \hspace{1cm} \text{and} \hspace{1cm} $$s_y = \frac{s_{yt}}{\sqrt{R_{yy}}}$$  (7.6b) \hspace{1cm} \\

Entering Equations 7.6a and 7.6b into the denominator of Equation 7.5 and then rearranging, we find that

$$r_{x,y} = \frac{c_{x,y}}{s_x s_y} \frac{s_x}{\sqrt{R_{xx}}} \frac{s_y}{\sqrt{R_{yy}}}$$

$$= \frac{c_{x,y}}{s_x s_y \sqrt{R_{xx} R_{yy}}}.$$
And again, we realize that a correlation is equal to a covariance divided by standard deviations (again, see Chapter 3’s discussion of the correlation coefficient, Equation 3.5). In this case, we divide the covariance between two sets of true scores (i.e., $c_{XY}$) by the standard deviations of those true scores (i.e., $s_X$ and $s_Y$), producing the correlation between true scores ($r_{XY}$). This simplifies the equation to

$$r_{XY} = \frac{r_{XX} R_{YY}}{R_{YY}}.$$ 

This brings us back to Equation 7.3. Thus, CTT implies directly that the correlation between two measures (i.e., between observed scores) is determined by the correlation between psychological constructs and by the reliabilities of the measures.

To illustrate this, imagine that we wish to examine the association between self-esteem and academic achievement. We conduct a study in which participants complete a self-esteem questionnaire and a measure of academic achievement. Imagine that the true correlation between the constructs is .40 (i.e., $r_{XY} = .40$). Of course, we would not actually know this true correlation; in fact, the entire point of conducting a study is to uncover or estimate this correlation.

In addition, imagine that both measures have good reliability—say, reliability is .80 for the self-esteem questionnaire and .86 for the academic achievement test. According to Equation 7.3, the correlation between the two measures will be

$$r_{XY} = r_{XY} \sqrt{R_{XX} R_{YY}},$$

$$= .40 \sqrt{.80 \times .86},$$

$$= .40 \times .829,$$

$$= .33.$$ 

Note that the correlation between observed scores on the two measures is smaller than the correlation between the two constructs. Specifically, the correlation between the two constructs is .40, but the correlation that we would actually obtain in our study is only .33. This discrepancy is a result of measurement error, as we will explain next.

**Measurement Error (Low Reliability) Attenuates the Observed Associations Between Measures**

The discrepancy between observed associations and true associations reflects four important implications of Equation 7.3 (again, if the assumptions of CTT hold true, see Loken & Gelman, 2017; Nimon et al., 2012). In this section, we describe and illustrate these important implications.

First, in research, observed associations (i.e., between measures) will always be weaker than true associations (i.e., between psychological constructs). This arises from two facts of life in measurement. One is that measurement is never perfect. Although scientists might develop very precise measures of their constructs, measures will always be affected by measurement error to some degree. That is, measures are not perfectly reliable. A second fact of life in measurement is that
imperfect measurement weakens or “attenuates” observed associations. For example, as shown by Equation 7.3, any time that reliabilities are less than perfect, an observed correlation will be weaker (i.e., closer to 0) than the true correlation. For example, what would the observed correlation be if the true correlation is .40 and both measures were nearly perfectly reliable (say both had reliabilities of .98)?

\[ r_{xy} = 0.40(0.98)(0.98), \]
\[ = 0.40(0.98), \]
\[ = 0.39. \]

Thus, even the slightest imperfections in measurement will begin to attenuate observed associations. In sum, given that measurement is never perfect and that imperfect measurement attenuates our observed associations, our observed associations will always be weaker than the true associations.

A second important implication of Equation 7.3 is that the degree of attenuation is determined by the reliabilities of the measures. Simply put, the poorer the measure, the greater the attenuation. More precisely, measures that have low reliability produce more extreme attenuation than measures that have high reliability. Consider again our example of the association between self-esteem and academic achievement, in which we assumed that the true correlation was .40. We saw earlier that, using measures with reliabilities of .86 and .80, the correlation between measures was attenuated to .33. What would the correlation be if the measures of self-esteem and academic achievement were poorer? For example, if the reliabilities were only .60 for the self-esteem measure and .50 for the academic achievement measure, then we would obtain a correlation of only .22:

\[ r_{xy} = 0.40(0.60)(0.50), \]
\[ = 0.40(0.30), \]
\[ = 0.22. \]

Obviously, this is a more extreme discrepancy between the true correlation and the observed correlation than we saw in the earlier example. Furthermore, the observed correlation can be extremely attenuated even if only one of the measures has poor reliability. For example, imagine that the academic achievement measure has good reliability (say .80) but the self-esteem questionnaire has very poor reliability (say .30). In this case, the observed correlation is attenuated to .20:

\[ r_{xy} = 0.40(0.80)(0.30), \]
\[ = 0.40(0.24), \]
\[ = 0.20. \]

In sum, the degree of attenuation is determined by the reliabilities of the two measures. If even one measure has poor reliability, the observed correlation can be considerably weaker than the true correlation. As we will see, such attenuation can have important effects on the accuracy with which we interpret research findings.

A third important implication of the fact that measurement error attenuates associations is that error constrains the maximum association that could be found...
between two measures. For example, imagine that you are interested in the association between academic motivation and academic achievement. You hypothesize that students who have relatively high levels of academic motivation will have relatively high levels of academic achievement. That is, students who care strongly about doing well in school should generally perform better than students who do not care about doing well (presumably because highly motivated students are more inclined to do homework, to pay attention in class, etc.). Although you believe that your hypothesis is reasonable, you do not know the size of the association between motivation and achievement; in fact, you do not even know if there truly is any association between the constructs. Therefore, you conduct a study in which your participants complete a measure of academic achievement and a measure of academic motivation.

While planning your study, you search for measures of your two constructs, and you pay careful attention to the reliabilities of the various measures that you might use. You find a highly reliable measure of academic achievement (say reliability = .86), but the only available measure of academic motivation has a poor reliability (say .40). Because you are familiar with Equation 7.3 and you know that measurement error attenuates the correlation between measures, you rightfully worry about the poor reliability of the motivation measure. You might even wonder about the highest possible correlation that you might obtain. That is, if your hypothesis is exactly correct and there is a perfect association between motivation and achievement, then what would your study reveal? By using Equation 7.3 and assuming a perfect association between the constructs (i.e., assuming that $r_{XY} = 1.0$), you find that

$$r_{XY} = 1.00 \sqrt{(0.86)(0.40)} = 1.00(0.587) = 0.59.$$  

This simple analysis tells you that even if your hypothesis is completely accurate and motivation is perfectly correlated with achievement, your study would reveal a correlation of “only” .59 between the two measures. Although a correlation of .59 would probably be taken as reasonable support for your theory, you should realize that this value represents the maximum possible correlation that you could hope to obtain if you use the two measures that you have chosen. That is, given the reliabilities of your two measures, you can obtain a correlation of .59 at best.

This information can be useful when you interpret the correlation that you actually obtain in your study. Because motivation and achievement are probably not perfectly correlated (i.e., it is likely that $r_{XY} < 1.0$), you will probably obtain a correlation that is quite a bit weaker than .59. In fact, you are likely to obtain a correlation much closer to .30 or even weaker, which might lead you to conclude that motivation is only moderately or even weakly associated with achievement. However, it might be very useful to interpret your results in the context of the best possible results that you could have hoped to obtain given the limits of your measures. Indeed, a correlation of .30 is much more compelling when you realize that a correlation of .59 was the best you could have hoped for considering the reliability of your measures.
A fourth important implication of Equation 7.3 is that it is possible to estimate the true association between a pair of constructs. When researchers actually conduct a study, they know or can estimate all but one component of Equation 7.3. Specifically, they do not know the true correlation between constructs; however, they can compute the observed correlation between the measures, and they can estimate the measures’ reliabilities (e.g., using a procedure discussed in Chapter 6).

Knowing all but one component of Equation 7.3, researchers can solve for the unknown component. In fact, the equation can be rearranged algebraically, producing a way of estimating the true correlation:

\[
\hat{r}_{XY} = \frac{r_{XY}}{\sqrt{R_{XX} R_{YY}}} \tag{7.7}
\]

Equation 7.7 is known as the correction for attenuation because it allows researchers to estimate the correlation that would be obtained if it were not affected by attenuation. That is, it allows researchers to estimate the true correlation—the correlation that would be obtained if perfectly reliable measures had been used in the study. If the measures were perfectly reliable, then the observed correlation would be exactly equal to the correlation between true scores.

As an illustration, assume that your study of the association between academic achievement and academic motivation revealed an observed correlation of \( r_{XY} = .26 \) based on the motivation questionnaire with an estimated reliability of .40 and the achievement test with an estimated reliability of .86. Of course, you do not know the true correlation, but you can use Equation 7.7 to estimate it:

\[
\hat{r}_{XY} = \frac{.26}{\sqrt{.40 \times .86}} = \frac{.26}{.587} = .44.
\]

Thus, if all the assumptions of CTT are correct (e.g., error affects test scores as if it is random), then you estimate that the true correlation between motivation and achievement is .44.

The correction for attenuation is an important perspective within the overall connections among reliability, measurement error, observed associations, and true associations; however, the correction procedure is not used explicitly very often in real research. That is, when reading research reports, you do not often see researchers conducting the correction for attenuation. Interestingly, recent developments in statistical analyses conduct an implicit correction for attenuation. Some of you might be familiar with a statistical procedure called structural equation modeling or latent variable modeling. Briefly, this procedure is designed (in part) to estimate the associations among unobservable psychological constructs by separating them...
Reliability, Effect Sizes, and Statistical Significance

The fact that measurement error (i.e., low reliability) attenuates observed associations has several implications for interpreting and conducting research. First, the results of a study should always be interpreted in the context of reliability. Although we have been discussing the “results” in terms of the observed correlation between measures, there are several different kinds of results that you might see. At least two basic kinds of results should be of interest in behavioral research.

These results—effect sizes and statistical significance—are affected heavily by reliability and measurement error. People who read and/or produce research should recognize these effects and take them into account when considering the results of scientific research.

Effect Sizes. Effect sizes are values that represent the results of a study as a matter of degree. For example, some effect sizes, such as the correlation coefficient, reflect the degree of association among variables, and others reflect the size of the differences among groups or conditions. Indeed, the previous sections describe the way in which one particular effect size (i.e., the correlation between observed scores on two variables) is affected by reliability. In addition to correlation coefficients, effect sizes also include statistics such as regression coefficients, $R^2$ values, $\eta^2$ values (from analysis of variance), odds ratios, and Cohen’s $d$ (from $t$ tests of means).

More and more, researchers are recognizing that effect sizes are a crucial part of their scientific results, arguably the most crucial. To fully understand the nature of their scientific findings, researchers should compute and interpret one or more effect sizes (Cumming, 2014; Wilkinson & APA Task Force on Statistical Inference, 1999). In fact, some researchers have suggested that “the primary product of a research inquiry is one or more measures of effect size” (Cohen, 1990, p. 1310), and there is a clear trend for scholarly journals to require or encourage researchers to present effect sizes. Thus, it is crucially important to realize that effect sizes are affected directly by measurement error and reliability.

Although a full examination of such statistics is beyond the scope of this book, Table 7.1 summarizes the link between reliability and effect sizes (i.e., associations/differences) for three effect sizes that are very common in behavioral research—correlations, Cohen’s $d$, and $\eta^2$. These effect sizes reflect three fundamental types of analytic contexts: (1) the correlation is usually used to represent the association between two continuous variables (e.g., Intelligence and Academic Achievement), (2) Cohen’s $d$ is usually used when examining the association between a dichotomous variable and a continuous variable (e.g., Biological Sex and Academic Achievement), and (3) $\eta^2$ is usually used when examining the association between a categorical variable with more than two levels (e.g., Dosage of Medication: 0, 10, and 20 mg) and a continuous variable (e.g., Level of Depressive Affect).
For example, researchers examining sex differences in academic achievement might compute a Cohen’s $d$ to reflect the magnitude of those observed differences:

$$d_{X_0} = \frac{|\bar{X}_{o1} - \bar{X}_{o2}|}{\sqrt{s^2_{o1} + s^2_{o2}} / 2}$$

In this particular form of the equation (which is appropriate when the two groups have equal numbers of participants), $\bar{X}_{o1}$ and $\bar{X}_{o2}$ are the two groups’ observed mean levels of achievement, and $s^2_{o1}$ and $s^2_{o2}$ are the two groups’ variances of observed achievement scores. Cohen’s $d$ is interpreted as the difference between two groups, in terms of the number of standard deviation units on the dependent variable. The lower limit of Cohen’s $d$ is 0 (reflecting no difference between the two groups’ mean levels of achievement), but its upper limit is, in theory, unlimited. Usually, the values fall between 0 and 1.5 (e.g., two groups are 1.5 standard deviations apart), with larger values reflecting bigger differences between the groups’ means. Table 7.1 shows that the observed value for Cohen’s $d$ ($d_{X_0}$) depends on two things: (1) the true value of Cohen’s $d$ (i.e., $d_{X_0}$, the degree to which the male and female participants differ in their true average levels of academic achievement) and (2) the reliability of the measure of academic achievement (i.e., $R_{XX}$).
The hypothetical data in Table 7.2 illustrate this effect for Cohen’s d. This dataset is much smaller than is typically found (or recommended) in behavioral research, but it reflects a hypothetical set of males and females who are measured on academic achievement (using a 0–4 scale). The “Observed Score” column presents these measured scores. If we temporarily pretend to be omniscient, then let’s say we also know the participants’ true levels of academic achievement and the degree to which their observed scores are affected by measurement error. Within each of these groups, the reliability of scores on the DV (i.e., Academic Achievement) is much poorer than we would typically like, being only $R_{xx} = .49$. Note that the Cohen’s d value for the true scores is extremely robust ($d_{X_T} = 1.52$):

$$d_{X_T} = \frac{|3.025 - 2.375|}{\sqrt{0.182 + 0.182}} = 0.650 \div 0.427 = 1.52.$$ 

This “true score effect size” value indicates that the “true” means are approximately 1.5 standard deviations apart—an extremely large difference, suggesting that the females truly are much more academically capable than males. In contrast, the Cohen’s d value for the observed scores is noticeably less, $d_{X_o} = 1.07$. This “observed score effect size” is consistent with the equation in Table 7.2, as $1.07 = 1.52 \times 0.49$. As discussed in Chapter 5, measurement error creates larger variances among the observed scores ($\sigma_{X_o}^2 = \sigma_{X_T}^2 = 0.369$) than among true scores ($\sigma_{X_T}^2 = \sigma_{X_T}^2 = 0.182$). Moreover, the relatively large variance among the observed scores reduces (i.e., attenuates) the observed effect size compared with the true effect size. Thus, researchers who interpret only the observed effect size will fail to understand the true psychological results of their study, underestimating the effect by a robust amount.

In sum, reliability affects many kinds of effect sizes, with good reliability producing better estimates of true effect sizes. All else being equal, better reliability produces larger observed effect sizes, while poorer reliability attenuates the observed effect sizes.

**Statistical Significance.** A second important kind of result in behavioral research is statistical significance, which, roughly speaking, is related to a researcher’s confidence in a result. That is, if a result is statistically significant, then researchers generally interpret it as being a “real” finding and not simply a fluke. As you might imagine, researchers typically hope that their research produces findings that are statistically significant.

Again, a full examination of such issues is beyond the scope of this book; however, it is important to realize that statistical significance is affected strongly by the size of the observed effect in a study (e.g., the size of an observed correlation or of an observed Cohen’s d value). All else being equal, larger observed effect sizes make it more likely that a result will be statistically significant.

Thus, through its impact on effect sizes, reliability indirectly affects statistical significance—higher reliability allows for higher observed effect sizes, which increases the likelihood that a result will be statistically significant. Conversely, low
reliability might contribute to a lack of statistical significance—lower reliability attenuates observed effect sizes, which decreases the likelihood that a result will be statistically significant.

This effect is again presented in Table 7.1 (in the “Inferential Statistic” column) and illustrated in the hypothetical data in Table 7.2. For example, Table 7.2 shows that the independent groups $t$ test of the true scores is significant ($t_{(10)} = 2.41, p = .04$). That is, the true psychological “story” in Table 7.2 is that males and females differ significantly in terms of their true levels of academic achievement. However, the independent groups $t$ test of the observed scores is not statistically significant.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Observed Score ($X_o$)</th>
<th>True Score ($X_t$)</th>
<th>Measurement Error ($X_e$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.6</td>
<td>2</td>
<td>$-0.4$</td>
</tr>
<tr>
<td>2</td>
<td>2.05</td>
<td>2.25</td>
<td>$-0.2$</td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
<td>2.5</td>
<td>$-0.2$</td>
</tr>
<tr>
<td>4</td>
<td>2.35</td>
<td>1.75</td>
<td>$0.6$</td>
</tr>
<tr>
<td>5</td>
<td>2.35</td>
<td>2.75</td>
<td>$-0.4$</td>
</tr>
<tr>
<td>6</td>
<td>3.6</td>
<td>3</td>
<td>$0.6$</td>
</tr>
<tr>
<td>Females</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.25</td>
<td>2.65</td>
<td>$-0.4$</td>
</tr>
<tr>
<td>8</td>
<td>2.7</td>
<td>2.9</td>
<td>$-0.2$</td>
</tr>
<tr>
<td>9</td>
<td>2.95</td>
<td>3.15</td>
<td>$-0.2$</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2.4</td>
<td>$0.6$</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>3.4</td>
<td>$-0.4$</td>
</tr>
<tr>
<td>12</td>
<td>4.25</td>
<td>3.65</td>
<td>$0.6$</td>
</tr>
</tbody>
</table>

| Means (variance) | |
|------------------|------------------|------------------|
| Group 1          | 2.375 (0.369)    | 2.375 (0.182)    | 0 (0.187)                |
| Group 2          | 3.025 (0.369)    | 3.025 (0.182)    | 0 (0.187)                |
| Cohen’s $d$      | 1.07             | 1.52             |
| $t$ value        | 1.69             | 2.41             |
| $p$ value        | .12              | .04              |
Thus, according to the observed scores, males and females do not appear to differ in terms of their academic achievement. Of course, researchers have access only to the observed data, not to true scores.

Thus, in the example in Table 7.2, the observed data would lead researchers to inaccurate conclusions about the effect of the independent variable (i.e., Sex) on the DV. As illustrated in Table 7.2, this inaccurate conclusion is driven (in part) by the poor reliability of the observed scores on the DV.

In sum, reliability is important in part because it has a clear and robust effect on two key results in a typical scientific study. By affecting effect sizes and statistical significance, reliability can have a fundamental impact on the results that researchers (and readers of research) see and interpret. If poor reliability biases these results, then researchers can be misled into making inaccurate conclusions about their work. Therefore, it is important that effect sizes and statistical significance be interpreted with close attention to the reliability of the measures used in the study. Measures with poor reliability are likely to underestimate the true effect sizes and are thus relatively likely to produce nonsignificant results.

Implications for Conducting and Interpreting Behavioral Research

The effects of reliability on effect sizes and on statistical significance are vital issues when interpreting the results of a study. There are at least three important implications of considering reliability when drawing psychological conclusions from research.

The first important implication is that researchers (and readers of research) should always consider reliability’s effects on their results when interpreting effect sizes and/or statistical significance. Imagine that you are a member of a school board that’s interested in enhancing students’ academic achievement. The board is considering two possible programs that are designed to enhance achievement. One is based on a theory that self-esteem affects academic achievement—students who feel good about themselves will perform better in school. Therefore, one program would be designed to increase students’ self-esteem, which, in turn, could have beneficial effects on their academic achievement. The second potential program is based on a theory that academic motivation affects academic achievement—students who are properly motivated will perform better in school. This program would be designed to increase students’ academic motivation, which could have beneficial effects on their achievement. Unfortunately, the school district has enough money to fund only one program, and the board wants to fund the program that might make the biggest impact on the students’ achievement.

A developmental psychologist at a local university agrees to conduct a study to determine which program might be most effective. Specifically, he will recruit a sample of students and measure all three constructs—academic achievement, self-esteem, and academic motivation. To keep our example simple, let us imagine that the researcher will compute two correlations: (1) the correlation between self-esteem and academic achievement and (2) the correlation between academic
motivation and academic achievement. The school board will fund the program for the variable that is most strongly associated with achievement, based on the assumption that it will have the larger impact on achievement. Therefore, if self-esteem is more strongly correlated with achievement, then the school board will fund the self-esteem program. However, if motivation is more strongly associated with achievement, then the school board will fund the motivation program.

The researcher collects the data and finds that the correlation between self-esteem and achievement ($r = .33$) is somewhat higher than the correlation between motivation and achievement ($r = .26$). Consequently, the school board begins to decide to fund the self-esteem program. However, you pause to ask the researcher about the reliability of his three measures. Although the researcher is surprised at the sophisticated level of your question, he tells you that the measure of achievement had an estimated reliability of .86, the measure of self-esteem had an estimated reliability of .80, and the measure of motivation had an estimated reliability of .40. What do you think of this psychometric information? Does it affect your opinion about which program should be funded? It should.

Take a moment to consider the fact that the self-esteem questionnaire seems to be more reliable than the motivation questionnaire. As we have discussed, all else being equal, higher reliability will produce higher observed correlations. But notice that the correlation involving motivation ($r = .26$) was only a bit smaller than the correlation involving self-esteem ($r = .33$), even though the motivation measure was substantially less reliable (reliability $= .40$) than the self-esteem measure (reliability $= .80$). Based on our discussion of attenuation, you should have a sense that the correlation involving motivation is attenuated to a much greater extent than is the correlation involving self-esteem. That is, you should begin to think that the observed correlation involving motivation is much lower than its true correlation, in comparison with the observed correlation involving self-esteem. In fact, you could correct both correlations for attenuation by using Equation 7.7 (see alternative procedures in Charles, 2005, and Padilla & Veprinsky, 2012):

$$r_{xy} = \frac{r_{xy}}{\sqrt{R_{xx} R_{yy}}}.$$ 

The “corrected” correlation between motivation and achievement is

$$r_{x,y} = \frac{.26}{\sqrt{.86}(.40)} = .44.$$

The “corrected” correlation between self-esteem and achievement is

$$r_{x,y} = \frac{.33}{\sqrt{.86}(.80)} = .40.$$

These simple analyses reveal a finding with potentially important implications for the school board. Once you correct for attenuation, you see that the true (i.e., “corrected”) correlation involving motivation is actually somewhat higher than the true correlation involving self-esteem. That is, if the assumptions of CTT are
correct in this case, then motivation is somewhat more strongly related to achievement than is self-esteem. Based on this finding, the school board might reverse its initial decision and fund the motivation program instead of the self-esteem program.

Hopefully, this example illustrates the need to interpret the results of research in the context of reliability. If those of us who read research or conduct research fail to consider the effects of reliability and measurement error, then we risk misinterpreting results and reaching (or believing) faulty conclusions. This issue might be particularly important when two or more analyses are being contrasted with each other, as in the example of the school board. Two or more analyses will differ in terms of the constructs involved and in terms of the measures of those constructs. If the difference in measurement is ignored, then any observed difference in the results of the analyses might be mistakenly interpreted in terms of the difference in constructs. Thus, one important implication of our discussion of reliability is that the effects of reliability should always be considered when interpreting the results of research.

A second important research-based implication of our discussion is that researchers should try to use highly reliable measures in their work. Attenuation cannot be avoided altogether, because measurement is never perfect. However, the problem of attenuation can be minimized if researchers use highly reliable measures. By using measures that are highly reliable, researchers can be fairly confident that the observed associations between their measures are reasonably close approximations of the true associations between the constructs of interest.

Despite the advantages of using highly reliable measures, there are at least two reasons why researchers might use measures with poor reliability. One is that there might be no highly reliable measure of the construct of interest. In such a case, a researcher must decide between proceeding with a low-reliability measure or spending time and effort attempting to develop a highly reliable measure. Of course, there is no guarantee that the time and effort will produce a highly reliable measure, so this option may seem like a risky choice. A second reason why researchers might use measures with poor reliability is that they simply have not devoted sufficient effort to finding a reliable measure. In psychology, there are thousands of measures of all kinds of constructs, and these measures are sometimes difficult to identify and obtain. Some measures are published and easily available. Other measures are published but are copyrighted and require money and specific credentials to use. Still other measures are used in the research literature but are not described in enough detail for other researchers to use. Thus, a researcher who wishes to use a highly reliable measure of a specific construct can face a daunting task of identifying which measures are available and which seem to be the most reliable. In addition, he or she will need to obtain the measure (or measures) that seem to fit his or her needs most closely. Although this can be a simple process at times, at other times it can require money, effort, and a great deal of patience. Researchers must decide if the potential costs of identifying and obtaining highly reliable measures are worth the potential benefits, as we have described in this section. In most cases, they are.

A third research-based implication of the fact that reliability affects observed correlations is that researchers should report reliability estimates of their measures.
Above, we argued for the importance of interpreting research results in the context of reliability. However, readers can do this only if writers provide the relevant information. Thus, if you conduct research and prepare a report such as a thesis, dissertation, or manuscript to submit for publication, then you should include reliability estimates. As discussed in the previous chapter, estimates of reliability (e.g., coefficient alpha) can be obtained easily from most of the popular statistical software packages (e.g., SPSS, SAS). In many research reports, reliability estimates are provided along with other basic descriptive statistics, such as means and standard deviations. As a writer, you should be sure to include this information. As a reader, you should expect to find and think about this information (hopefully, the writer has as well!). If you find yourself reading a research report that fails to provide reliability information, then you should feel comfortable in contacting the author of the report and requesting the relevant information.

In sum, test reliability has important effects on behavioral research. Along with the true correlation between psychological constructs, reliability affects the observed association between measures. Although researchers should strive to use the most reliable measures available, they cannot or do not always do so. Consequently, a lack of reliability weakens or attenuates the results of their statistical analyses, potentially leading to misinterpretations of their findings. Along with those who conduct research, those who read research also should consider the attenuating effects of imperfect measurement when interpreting the results of behavioral research.

Test Construction and Refinement

The previous two sections have described some of the important ways in which reliability and measurement error affect research and practice in behavioral science. It should be clear that high reliability is a desirable quality of any psychological test or measurement. Thus, reliability is an important facet of test construction and refinement. In this section, we present some of the ways in which item information is evaluated in this process, and we highlight the role that reliability often plays.

As we saw in the previous chapter, internal consistency reliability is affected by two factors—test length and the consistency among the parts of a test. All else being equal, a longer test will be more reliable than a shorter test, and a test with greater internal consistency will be more reliable than a test with lower internal consistency.

In the test construction and refinement process, great attention is paid to the consistency among the parts of a test—typically in terms of the test items themselves. That is, test developers often examine various statistical characteristics of a test’s items. They do so to identify items that should be removed from the test or to find items that should be revised to enhance their contribution to the test’s psychometric quality. In general, items that enhance a test’s internal consistency are preferable to items that detract from the test’s internal consistency.

We will discuss three interconnected item characteristics that are important considerations in test construction and refinement: item means (or item difficulty), item variances, and item discrimination. In terms of reliability, the overarching
issue is item discrimination, which, as we shall see, is closely connected to internal consistency. Thus, our discussion will address the way in which the three item characteristics affect and reflect an item’s contribution (or lack thereof) to internal consistency. We will revisit the concepts of item difficulty and item discrimination in Chapter 14 (“Item Response Theory and Rasch Models”).

It is important to note that the procedures and concepts that we describe in this section should be conducted for each dimension being assessed by a test. As described in our earlier discussion of test dimensionality (Chapter 4), psychometric analysis should be conducted for each score that is produced by a test, with a score representing each psychological dimension underlying the responses to the test’s items. So for a unidimensional test, we would conduct the following analyses on all of the test’s items together as a single group (because all items are ostensibly combined together to create a single test score). However, for a multidimensional test, we would conduct the following analyses separately for each of the test’s dimensions. For example, imagine that a self-esteem test included 20 items, with Items 1 to 10 ostensibly reflecting social self-esteem and Items 11 to 20 ostensibly reflecting academic self-esteem. At a minimum, we would conduct the following analyses once for Items 1 to 10 and then again for Items 11 to 20.

To illustrate the psychometric examination of item means, variances, and discrimination, we will use the hypothetical data presented in Table 7.3. These data represent the responses of 10 people to a unidimensional test that includes five binary items in which a correct answer is coded “1” and an incorrect answer is coded “0.” Because it includes multiple items, the total test score is a composite variable, and a test developer might be concerned about evaluating and improving the psychometric quality of the test.

<table>
<thead>
<tr>
<th>Respondent</th>
<th>Item</th>
<th>Total</th>
<th>Excluding Item 1</th>
<th>Excluding Item 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maria</td>
<td>1 1 1 1 1</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Demetrius</td>
<td>1 1 1 1 1</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Rohit</td>
<td>1 1 0 0 1</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>James</td>
<td>1 0 1 1 1</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Antonio</td>
<td>0 0 1 0 1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Esteban</td>
<td>0 1 0 1 1</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Zoe</td>
<td>0 1 1 0 1</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Emory</td>
<td>1 0 0 0 0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Fitz</td>
<td>1 0 0 0 0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Claudette</td>
<td>0 0 0 0 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7.3 Example Data for Test Construction and Refinement
Using the Reliability Analysis procedure in the statistical package SPSS, we obtained a set of results that will help us evaluate the psychometric quality of the test (see the output in Table 7.4). The top section of this table reveals that the five-item test has an estimated reliability of only .59 (using coefficient alpha). For reasons discussed earlier in this chapter, we would prefer to have a test with greater reliability. Thus, we might wish to refine the test in a way that could improve its reliability for future use. The results of our SPSS analyses will help us examine the degree to which each item enhances or detracts from the test's quality, and this information can guide any test refinements.

**Item Discrimination and Other Information Regarding Internal Consistency**

As we have seen, one key to internal consistency reliability is the degree to which a test's items are consistent with each other. More specifically, internal consistency is the degree to which differences among persons' responses to one item are consistent with differences among their responses to other items on the test.

Thus, a test's internal consistency is intrinsically linked to the correlations among its items. For any particular item, its correlations with the other items reflect its consistency with those other items (and thus with the test as a whole). For example, if we find that an item is relatively strongly correlated with the other items on a test, then we know that the item is generally consistent with the other items. Consequently, we would know that the item enhances the internal consistency of the test. In contrast, if we find that an item is relatively weakly correlated with the other items on a test, then we know that the item is generally inconsistent with the other items. Consequently, we would suspect that the item reduces the internal consistency of the test.

With these considerations in mind, one important task is to determine which items contribute well to reliability and which detract from the test's reliability. A quick look at the correlation among a test's items might be very revealing. Indeed, a reliability-based test construction or refinement process might include an examination of the correlations among all of the items on a test.

For example, Table 7.4 presents these correlations in the “Interitem Correlation Matrix” output from the SPSS reliability analysis of the test responses in Table 7.3. A glance at these correlations reveals some good news and some bad news about the five-item test. The good news is that four items are relatively well correlated with each other. Specifically, Items 2 to 5 are generally intercorrelated with each other at levels of $r = .40$ or .50. Interitem correlations of this size indicate reasonable levels of internal consistency. The bad news is that one of the items is potentially problematic. Notice that Item 1 is totally uncorrelated with Item 2 and Item 3, only weakly correlated with Item 4 ($r = .25$), and negatively correlated with Item 5. These correlations suggest that Item 1 is not consistent with most of the other items on the test. The overall pattern of interitem correlations suggests that Items 2 through 5 are consistent with each other but that Item 1 needs to be revised or dropped from the test.
Table 7.4  SPSS Output From the Reliability Analysis of Data in Table 7.3

<table>
<thead>
<tr>
<th>Reliability Statistics</th>
<th>Cronbach’s Alpha</th>
<th>Cronbach’s Alpha Based on Standardized Items</th>
<th>Number of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.590</td>
<td>.594</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interitem Correlation Matrix</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>1.000</td>
<td>.000</td>
<td>.000</td>
<td>.250</td>
<td>-.408</td>
</tr>
<tr>
<td>Item 2</td>
<td>.000</td>
<td>1.000</td>
<td>.200</td>
<td>.408</td>
<td>.500</td>
</tr>
<tr>
<td>Item 3</td>
<td>0.000</td>
<td>.200</td>
<td>1.000</td>
<td>.408</td>
<td>.500</td>
</tr>
<tr>
<td>Item 4</td>
<td>.250</td>
<td>.408</td>
<td>.408</td>
<td>1.000</td>
<td>.408</td>
</tr>
<tr>
<td>Item 5</td>
<td>-.408</td>
<td>.500</td>
<td>.500</td>
<td>.408</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item–Total Statistics</th>
<th>Scale Mean If Item Deleted</th>
<th>Scale Variance If Item Deleted</th>
<th>Correlated Item–Total Correlation</th>
<th>Squared Multiple Correlation</th>
<th>Cronbach’s Alpha If Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>2.20</td>
<td>2.178</td>
<td>-.029</td>
<td>.410</td>
<td>.721</td>
</tr>
<tr>
<td>Item 2</td>
<td>2.30</td>
<td>1.567</td>
<td>.421</td>
<td>.337</td>
<td>.492</td>
</tr>
<tr>
<td>Item 3</td>
<td>2.30</td>
<td>1.567</td>
<td>.421</td>
<td>.337</td>
<td>.492</td>
</tr>
<tr>
<td>Item 4</td>
<td>2.40</td>
<td>1.378</td>
<td>.623</td>
<td>.410</td>
<td>.366</td>
</tr>
<tr>
<td>Item 5</td>
<td>2.00</td>
<td>1.778</td>
<td>.395</td>
<td>.627</td>
<td>.517</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item Statistics</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>.60</td>
<td>.516</td>
<td>10</td>
</tr>
<tr>
<td>Item 2</td>
<td>.50</td>
<td>.527</td>
<td>10</td>
</tr>
<tr>
<td>Item 3</td>
<td>.50</td>
<td>.527</td>
<td>10</td>
</tr>
<tr>
<td>Item 4</td>
<td>.40</td>
<td>.516</td>
<td>10</td>
</tr>
<tr>
<td>Item 5</td>
<td>.80</td>
<td>.422</td>
<td>10</td>
</tr>
</tbody>
</table>
Although the interitem correlations offer insight into the internal consistency of a test, there are more efficient ways of evaluating the consistency issue. The interitem correlations in our example are fairly straightforward—there are only a few items, and the pattern of correlations is arranged in a rather clear way. In reality, most test development/refinement situations might be much more complicated, with many more items and with a more complex pattern of correlations. Thus, an examination of a matrix of interitem correlations might be somewhat impractical with real data; fortunately, alternative methods exist.

Item discrimination is a common concept for evaluating the degree to which an item might affect a test's internal consistency. Briefly stated, item discrimination is the degree to which an item differentiates people who score high on the total test from those who score low on the total test. From the perspective of reliability, we prefer to have items that have high discrimination values over those that have low discrimination values.

There are various ways of operationalizing an item's discrimination, one of which is the item–total correlation. We can compute the total score on a test (see Table 7.3) and then compute the correlation between an item and this total test score. The resulting correlation is called an item–total correlation, and it represents the degree to which differences among persons' responses to the item are consistent with differences in their total test scores. A high item–total correlation indicates that the item is consistent with the test as a whole (which of course is a function of all of the items within the test), which is a desirable characteristic. In contrast, a low item–total correlation indicates that the item is inconsistent with the test as a whole, which would be an undesirable characteristic from the perspective of reliability.

To illustrate this concept, the SPSS output labeled “Item–Total Statistics” in Table 7.4 presents “corrected” item–total correlations, which are correlations between an item and a “corrected” total test score. The corrected item–total correlation for Item 1 is the correlation between responses to Item 1 and the sum of the other four items on the test. That is, the “corrected” total test score in the analysis of Item 1 is the total that is obtained by summing all of the items except Item 1 (see the “Total Excluding Item 1” column in Table 7.3). If we compute the correlation between the Item 1 values and the “Total Excluding Item 1” values, then we obtain a value of $r = -0.029$. This value tells us that Item 1 seems to be generally inconsistent with the responses to the other four items. To compute a corrected item–total correlation for each item, SPSS computes a different corrected total test score for each item. As we have seen, the corrected item–total correlation for Item 1 requires a corrected total test score that excludes Item 1. Similarly, the corrected item–total correlation for Item 2 would require a corrected total test score that excludes Item 2 and so on. As we see in the SPSS output, all of the corrected item–total correlations are positive values of reasonable size, except for Item 1. On the basis of these results, we should consider dropping or revising Item 1.

Another form of item discrimination is particularly applicable for items that are scored in a binary manner, as we have in Table 7.3. An item’s item discrimination index ($D$) compares the proportion of high test scorers who answered the item correctly with the proportion of low test scorers who answered the item correctly. To do this, we begin by identifying a specific percentage of people with the highest
total test scores (say all respondents who scored in the upper 30%) and the same percentage of people with the lowest total test scores (say all respondents who scored in the lowest 30%). For the data in Table 7.3, the top 30% group includes Maria, Demetrius, and James, and the bottom 30% group includes Emory, Fitz, and Claudette. To compute the item discrimination index for an item, we next calculate the proportion of people within each group who answered the item correctly (as designated by a “1” in Table 7.3). For Item 1, we see that all three people in the “top 30%” group answered the item correctly, for a proportion of $p_{\text{high}} = 1.0$. In contrast, we see that only two of the three people in the “bottom 30%” group answered the item correctly, for a proportion of $p_{\text{low}} = .66$. Finally, we compute the item discrimination index by calculating the difference between these two proportions (see Brennan, 1971, for a broader view of this):

$$D = p_{\text{high}} - p_{\text{low}}.$$  \hspace{1cm} (7.8)

For Item 1, this results in an item discrimination index of .33:

$$D = 1.0 - 0.66 = .33.$$
test if we were to drop each item from the test. For example, the “alpha if item deleted” value for Item 1 is .721. This indicates that, if we drop Item 1 but retain the other four items, the reliability of the resulting four-item test would be estimated at $\alpha = .72$. Note that this value is clearly larger than the reliability estimate for the entire five-item test, which is $\alpha = .59$, as mentioned earlier. With these two reliability estimates in mind, we see that dropping Item 1 would actually improve the internal consistency of the test from .59 to .72. Thus, we would seriously consider refining the test by dropping Item 1. Also notice that reliability would decrease if any of the other items were dropped from the test—the other four “alpha if item deleted” values are less than .59.

In sum, we have examined several interconnected kinds of information that reveal an item’s effect on test score reliability. For example, we saw that Item 1 had relatively low interitem correlations, which suggested that Item 1 is inconsistent with the other items in the test. We then saw that, although Item 1’s discrimination index was greater than 0, its corrected item–total correlation was very close to 0, which suggested that Item 1 is inconsistent with total test scores in general. Finally, we saw that the test’s reliability would likely increase if we removed Item 1 from the test, which is consistent with the previous results that demonstrated Item 1’s inconsistency with the other four items.

Considering the results that we have discussed thus far, we have a relatively clear idea of how we might improve the five-item test. Clearly, we are likely to retain Items 2, 3, 4, and 5 in the test refinement process— they are well correlated with each other, and dropping any one of them would reduce the test’s reliability. However, we are likely to either drop Item 1 altogether or examine the item (i.e., its content, its wording, its response options, etc.) to see if we can improve it. It is possible that the test could be improved substantially if we were able to revise Item 1 in a way that makes it more consistent with the other four items. If so, then we could include the revised Item 1 along with the other four items to produce a stronger five-item test.

In the next section, we address two additional item characteristics that are sometimes evaluated in a test refinement process. Our discussion will highlight the ways in which item difficulty (i.e., item means) and item variance are related to an item’s effect on test reliability.

**Item Difficulty (Mean) and Item Variance**

An item’s mean and variance are potentially important factors affecting its contribution to the psychometric quality of a test. From the perspective of reliability, an item’s mean and variance are important because they may be related to the degree to which the item might be consistent with the other items on a test. Consequently, they have potential implications for an item’s effect on test score reliability.

As we discussed in Chapter 3, a correlation reflects the degree to which variability within one variable is consistent with variability within another variable. Indeed, a correlation is highly dependent on variance. Specifically, the correlation between two variables is a transformation of the covariance between the two variables. In turn, the covariance between two variables hinges on the existence of variance within each variable. Simply put, if a variable (e.g., responses to a test item) has no variability, then it will not be correlated with any other variable.
Based on the intrinsic link between correlation and variance, an item’s variance has potential implications for characteristics such as its interitem correlations, its item–total correlation, and its “alpha if item deleted” value. Items with limited variability are less likely to have good correlational characteristics than are items with substantial variability. Indeed, items that all respondents answer in the same way (e.g., all respondents answer correctly, or all answer incorrectly) are poor items from the perspective of reliability.

The link between an item’s variability and its psychometric quality can be extended to the item’s mean. In some cases, an item’s mean tells us about the item’s variability. Most psychological tests have practical limits on the responses that people can provide. For example, in the test presented in Table 7.3, the maximum score on each item is 1 and the minimum is 0. This “ceiling” and “floor” constrain the total test scores, and they have implications for the link between item means and item variances and, consequently, for the values of the covariances and correlations among items.

For example, imagine that Item 1 (in Table 7.3) had a mean of 1.0—what would this imply about the item’s variability? Because the maximum value of an individual’s response is 1, there is only one way that Item 1 can have a mean equal to 1.0. Specifically, Item 1 will have a mean of 1.0 only if every respondent answers the item correctly. Similarly, Item 1 will have a mean of 0 only if every respondent answers the item incorrectly. It should be clear that if every respondent answers an item in the same way, then the item will have no variability. And as we have discussed, if an item has no variability, then it is a poor test item from a reliability perspective. Thus, items that have “extreme” means (i.e., either very high or very low) are likely to have limited variability, and thus they are relatively likely to have poor psychometric qualities.

An item’s mean is sometimes interpreted in terms of “difficulty.” For example, the mean of Item 5 is .80 (shown in Table 7.4), which tells us that 80% of the respondents answered the item correctly (because we coded a correct answer as “1” and an incorrect answer as “0”). With fully 80% of respondents answering correctly, this is a relatively easy item. In contrast, the mean of Item 4 is .40, which tells us that only 40% of the respondents answered the item correctly. Thus, Item 4 appears to be more difficult than Item 5.

For binary response items, such as those presented in Table 7.3, CTT suggests that we would like to have items with difficulties of approximately .50. As discussed in the “Binary Items” section of Chapter 3, this ensures that items will have maximal variability. In turn, this avoids the difficulties associated with low variability.

**Summary**

In this section, we explained how reliability and measurement error affect the results of behavioral research. We showed that the correlation between the two measures is determined by the correlation between the psychological constructs being measured and by the reliabilities of the measures. These two factors will combine to influence the interpretation of the results of empirical research findings.

Test score reliability will also play a role in test score interpretation. We showed how test scores will regress to the mean of a distribution of scores and how the size
of this regression will depend on score reliability. Reliability will influence the confidence intervals created around particular scores; reliable test scores will be associated with smaller intervals than will less reliable scores.

We also presented some of the ways in which test item information is evaluated, and we highlighted the role that reliability often plays in this type of evaluation. Three interconnected item characteristics that are important considerations in test construction and refinement—item means, item variances, and item discrimination—were discussed in detail.

Suggested Readings

For a discussion of alternative ways of computing confidence intervals around scores:

For a discussion of regression to the mean that differs from the standard understanding of the phenomenon:

For a short and well-crafted technical discussion of attenuation, see pp. 133–136 in

For a description of an approach to statistics that recognizes the importance of effect sizes:

For a recent perspective on methods of correcting for attenuation:

For a detailed discussion of calculating an item discrimination index: