05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 81

CHAPTER 5

Reliability

Conceptual Basis

Nurses attempt to measure the length of babies at birth and at regular intervals thereafter. If you have ever watched an attempt to measure the length of a baby, you will not be surprised to learn that this is a difficult task. Babies squirm around erratically, and they resist attempts to stretch them out to their full length. Such squirming creates difficulties for nurses who are attempting to obtain accurate measurements of babies' lengths. Furthermore, some babies are more compliant than others, which means that some are less likely to squirm around than others. Again, this creates differences among babies, in that some babies may be more likely to be measured accurately than others. These kinds of problems have led researchers (e.g., Johnson, Engstrom, & Gelhar, 1997; Johnson, Engstrom, Haney, & Mulcrone, 1999) to ask questions about the reliability of these measurements.

Imagine that a nurse is asked to measure the lengths of 10 different babies, and imagine that there was some way to know beforehand (but unknown to the nurse) each baby's true length. You could, in theory, compare each baby's measured length with his or her true length. Moreover, you could examine the differences among babies' measured lengths and compare them to the differences among babies' true lengths. Ideally, you would find good consistency between these two sets of differences. That is, you would hope to find that differences among the babies' measured lengths were consistent with differences in their actual lengths. If so, then you would conclude that the measurement procedure produced length values that were reliable.

Throughout this book, we have emphasized the importance of understanding psychological variability—psychological tests are useful only to the degree that they accurately reflect true psychological differences. Again, in a research context, behavioral science strives to quantify the degree to which differences in one variable (e.g., intelligence) are associated with differences in other variables (e.g., parenting styles, preschool experience, age, academic performance, aggression, gender, etc.). Psychological measures are used to assess and represent these differences. In an

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 82

applied context, practitioners strive to make decisions about people, and they use psychological measures to inform those decisions. Such decisions rest on the assumption that differences among people exist and have important implications. Thus, psychological measurement always hinges on the ability to reflect real psychological differences accurately. This ability is at the heart of reliability.

This chapter introduces classical test theory (CTT), which is a measurement theory that defines the conceptual basis of reliability and that outlines procedures for estimating the reliability of psychological measures (Gulliksen, 1950; Magnusson, 1967). For example, suppose that we give a burnout questionnaire to a group of people, and we found that people differ in their scores on the questionnaire. We would like to assume that differences in their questionnaire scores accurately reflect differences in their true levels of burnout. According to classical test theory, a test's reliability reflects the extent to which the differences in respondents' test scores are a function of their true psychological differences, as opposed to measurement error. We hope that all of our measures are highly reliable.

Although people sometimes speak as if reliability is an all-or-none property of the results of a measurement procedure, in fact, reliability is on a continuum. A procedure for measuring something will be more or less reliable.

Notice that reliability is itself a theoretical notion. Reliability is a feature, presumably, of the results of procedures for measuring characteristics of objects or psychological characteristics of people. Just as a psychological attribute such as intelligence is an unobserved feature of a person, reliability is an unobserved feature of test scores. Furthermore, just as we must estimate a person's level of intelligence, we must estimate a test's reliability. In this chapter, we will describe the theoretical basis of reliability from the perspective of classical test theory. In the next chapter, we will describe procedures for estimating a test's reliability. We will show that it is possible, given certain assumptions of CTT, to calculate numerical values that estimate the *degree* to which scores from a measure are or are not reliable.

Overview of Reliability and Classical Test Theory

According to CTT, reliability is a test property that derives from observed scores, true scores, and measurement error. We refer to values that are obtained from the measurement of some characteristic of a person as *observed scores.* In contrast, the real amounts of that characteristic are referred to as *true scores.* In our length example, a baby's length as determined by the nurse would be an observed score, and a baby's real length would be a true score. Ideally, test users would like to interpret individuals' observed scores as good estimates of their true scores because most behavioral research and decision making are intended to reflect respondents' true psychological characteristics.

Reliability reflects the extent to which differences in respondents' observed scores are consistent with differences in their true scores. More specifically, the reliability for a measurement procedure depends on the extent to which differences in respondents' observed scores on the measure can be attributed to differences in their true scores, as opposed to other, often unknown, test and test administration

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 83

characteristics. The extent to which these "other" characteristics contribute to differences in observed scores is referred to as *measurement error,* or just error, because they create inconsistency between observed scores and true scores.

When measuring the quantity of anything, including features of physical objects or psychological characteristics of people, the results of the measurement will always be unreliable to some extent. That is, there is no such thing as a perfectly reliable measure. In fact, even if such a measure existed in some Platonic world, we would not be about to identify it as perfectly reliable given the limitations of our empirical abilities.

It is usually impossible to know all of the sources of measurement error affecting test scores. In the case of measuring the length of babies, we can speculate that some of the error might be related to how much each baby squirms while being measured. Some babies are going to squirm more than others. If the measurements of length are affected by amount of squirming, then the accuracy of the length measures will be affected by the amount each baby squirms. Some babies' squirming might cause their nurses to underestimate their true length, but other babies' squirming might cause their nurses to overestimate their true length. The effects of squirming are considered measurement error because they detract from the accurate measurement of babies' true lengths. Other sources of error in the babies' measurement might include the fact that different nurses might record the measurements. If each baby is measured by a different nurse, and if some nurses are more careful in taking their measurements, then some babies will be more accurately measured than others. Differences in nurses' "measurement care" will obscure differences among babies' true lengths. There are many possible sources of error that might affect observed measurements, thereby obscuring the true differences among babies. Some of these sources of error might be subtle (e.g., nurses' carefulness), and some might be more obvious (e.g., squirming). There is no way to account for all of the possibly subtle factors that might affect observed scores.

Such errors also influence the measurement of psychological attributes. Consider what might happen if a class of schoolchildren takes a mathematics achievement test. We would like to think that a child's score on the test is an accurate estimate of the child's true knowledge of mathematics; however, it is clear that factors other than knowledge might influence the children's test performance. Some children who are taking the test might have a cold on the day that they take the test. The cold might make them groggy, which in turn causes them to perform worse on the test than they "truly" could perform, given their true mathematical ability. Some children might have eaten a nutritious breakfast, which helps them feel alert and energetic, thereby causing them to perform quite well on the test. Some children might happen to make many "lucky guesses" on the test, which makes their test score higher than it should really be, given their true mathematical ability. Some children might compute the math answers correctly but mistakenly circle the wrong choice on an answer sheet, producing test scores that artificially underestimate their "true" mathematical ability. Such temporary and transient factors—amount of sleep, emotional state, physical well-being, guessing, misrecording answers—could artificially inflate or deflate the children's test scores relative to their true scores. Each of these factors might be a source of measurement error compromising the quality of the test scores.

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 84

In order to evaluate the reliability of scores from any measure, we must estimate the extent to which individual differences in observed scores are a function of measurement error versus the extent to which they are a function of true or real score differences among respondents.

Observed Scores, True Scores, and Measurement Error

The degree of reliability for a test depends on two things: (a) the extent to which differences in test scores can be attributed to real inter- or intraindividual differences and (b) the extent to which such differences are a function of measurement error. In classical test theory, a person's observed score on a test is a function of that person's true score, plus error. If X_o represents an individual's observed test score, if X_t is the individual's true score on a psychological characteristics, and if X_e is the amount of error affecting the individual's responses, then we can write the following formula to represent this assumption:

$$X_o = X_t + X_e. \tag{5.1}$$

To illustrate this point, we have constructed an artificial data set representing six people's responses to a self-esteem questionnaire (see Table 5.1a). For the sake of this example, we will pretend that we know each person's true level of self-esteem (i.e., each person's true score, X_t). Of course, we would never actually know an individual's true score—this example is intended solely to explain the theoretical basis of reliability. From this "omniscient" perspective, we see that Ashley truly has the highest level of self-esteem in this sample (Ashley's $X_t = 130$), that Bob has the next highest level (Bob's $X_t = 120$), and so on.

In addition, we pretend that we know the degree to which each individual's questionnaire score is affected by measurement error. For example, Ashley happened to take the self-esteem questionnaire only an hour after learning that she had earned a D on a biology test. Because of this disappointing grade, she felt unusually bad about herself when she took the self-esteem questionnaire. Notice that Ashley's error score (Ashley's X_e) is –10, reflecting the fact that this event temporarily lowered her apparent self-esteem score. In contrast, Bob happened to take the test an hour after learning that he had been accepted into law school. Although Bob generally has a relatively high level of self-esteem (i.e., his true self-esteem score is relatively high compared to the rest of the sample), the good news about law school makes him feel even better about himself than he usually does. Notice that Bob's error score (Bob's X_e) is +25, reflecting the fact that this good news is temporarily raising his apparent self-esteem score.

As Table 5.1a shows, the respondents' observed scores on the self-esteem questionnaire are determined by their true levels of self-esteem and by the "error" effect of random events or states. For example, Ashley's observed score is as follows:

$$\begin{aligned}&\text{Ashley's } X_o = \text{Ashley's } X_t + \text{Ashley's } X_e,\\ &\text{Ashley's } X_o = 130 + (-10),\\ &\text{Ashley's } X_o = 120.\end{aligned}$$

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 85

Table 5.1 Responses to an Original Self-Esteem Questionnaire and a Revised Self-Esteem Questionnaire

(a) Responses to Original Self-Esteem Questionnaire

Respondent	(X_o) *Observed Score*		(X_t) *True Score*		(X_e) *Error*
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Standard Deviation	24.66		17.08		17.80
Reliability = R_{xx} =	.48	r_{ot} =	.69	r_{oe} =	.72
r_{te} =	.00	r_{ot}^2 =	.48	r_{oe}^2 =	.52

(b) Responses to Revised Self-Esteem Questionnaire

Respondent	(X_o) *Observed Score*		(X_t) *True Score*		(X_e) *Error*
Ashley	135	=	130	+	5
Bob	130	=	120	+	10
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	100	=	90	+	10
Felicia	85	=	80	+	5
Mean	105		105		0
Variance	408.33		291.67		116.67
Standard Deviation	20.21		17.08		10.80
Reliability = R_{xx} =	.71	r_{ot} =	.84	r_{oe} =	.53
r_{te} =	.00	r_{ot}^2 =	.71	r_{oe}^2 =	.29

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 86

Again, this "omniscient" example illustrates the first simple but fundamental theoretical assumption of classical test theory—that *observed scores on a psychological measure are determined by respondents' true scores and by measurement error.*

The CTT perspective on reliability makes a very important *assumption about measurement error.* Specifically, it assumes that error occurs as if it is random. In part, this means that measurement error is just as likely to inflate any particular score as it is to decrease any particular score. We assume that individuals' responses to a psychological test are affected in unpredictable ways that might make their observed scores artificially high or artificially low. Consider Ashley and Bob. It was simply chance that Ashley took the self-esteem questionnaire only an hour after hearing bad news, thereby lowering her observed score as compared to her true, stable level of self-esteem. Similarly, it was simply chance that Bob took the questionnaire after hearing good news, thereby raising his observed score as compared to his true, stable level of self-esteem. Across the entire sample of respondents, error artificially inflates some people's scores, and it artificially deflates other people's scores.

Because error affects scores as if it is random, the inflation and deflation caused by error is independent of the individuals' true levels of self-esteem. That is, measurement error can affect someone with a high true level of self-esteem in the same way (and to the same degree) as it affects someone with a low true level of self-esteem. Again, consider Ashley and Bob. The events that are temporarily affecting their responses have nothing to do with their true level of self-esteem. The timing of Ashley's grade on her biology test and the timing of Bob's news about law school are completely unrelated to how high or low their true levels of self-esteem are. The artificial data in Table 5.1a illustrate this general point. Notice that the size and direction (positive or negative) of the error effects are spread equally for respondents across the entire range of true scores. For each "high-esteem" person whose observed score is artificially deflated by measurement error, there is a high-esteem person whose observed score is artificially inflated. The same is true for people with low true levels of self-esteem.

There are two important consequences of this assumption about error. First, error tends to cancel itself out across respondents. That is, error inflates the scores of some respondents and deflates the scores of other respondents in such way that the average effect of error across respondents is zero. Indeed, Table 5.1a shows that the mean of the six error scores is exactly zero (i.e., $\overline{X}_e = 0$). The second consequence of the apparent randomness of error is that error scores are uncorrelated with true scores. As described above, error affects observed scores in ways that are independent of the respondents' true levels of self-esteem. Therefore, if we compute the correlation between individuals' true scores and their error scores in Table 5.1a, we find that the correlation is exactly zero (i.e., ($r_{te} = 0$). These two consequences have important implications for reliability, as we shall soon see.

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 87

Variances in Observed Scores, True Scores, and Error Scores

As mentioned earlier, reliability hinges on the degree to which *differences* in observed scores are consistent with *differences* in true scores. Put another way, reliability hinges on the links among observed score variability, true score variability, and error score variability. Given the importance of variability for interpreting and evaluating psychological measurement, we need to understand how the first assumption of classical test theory (i.e., that, for each individual, $X_o = X_t + X_e$) extends to the differences among people.

This extension might make the most sense if we begin by illustrating how the true differences between people can be obscured by differences in measurement error. Look at the individuals' true scores in Table 5.1a and focus on the difference between Ashley and Bob. Notice that Ashley's true score (Ashley's $X_t = 130$) is 10 points higher than Bob's (Bob's $X_t = 120$). However, notice that Ashley's observed score on the questionnaire (Ashley's $X_o = 120$) is 25 points lower than Bob's observed score (Bob's $X_o = 130$). Obviously, the difference between Ashley's and Bob's true scores is inconsistent with the difference between their observed scores—Ashley's true score is higher than Bob's true score, but her observed score is lower than Bob's observed score:

$$\text{Ashley's } X_t - \text{Bob's } X_t = 130 - 120 = +10$$

$$\text{Ashley's } X_o - \text{Bob's } X_o = 120 - 145 = -25$$

This inconsistency is created by the measurement error that artificially deflated Ashley's observed score but artificially inflated Bob's observed sore. Of course, this inconsistency means that the apparent 25-point difference between Ashley and Bob (on the self-esteem questionnaire) is a poor reflection of the real 10-point difference between Ashley and Bob (in their true, stable levels of self-esteem).

Because such inconsistencies potentially affect the differences among all the respondents, let us consider the relevant variances across all the participants. Variances for this hypothetical data set are computed in the standard way. For example, variance among the error scores (s_e^2) is based on using error scores (X_e) in the computations:

$$s_e^2 = \frac{\sum(X_e - \bar{X}_e)^2}{N}, \tag{5.2}$$

$$s_e^2 = \frac{(-10-0)^2 + (25-0)^2 + (-15-0)^2 + (-15-0)^2 + (25-0)^2 + (-10-0)^2}{6},$$

$$s_e^2 = \frac{(-10)^2 + (25)^2 + (-15)^2 + (-15)^2 + (25)^2 + (-10)^2}{6},$$

$$s_e^2 = \frac{100 + 625 + 225 + 225 + 625 + 100}{6},$$

$$s_e^2 = \frac{1900}{6},$$

$$s_e^2 = 316.67.$$

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 88

This value represents the degree to which error affected different people in different ways. Again, the fact that error affects people differently—artificially inflating some people's scores and artificially deflating other people's scores—is what obscures the true differences among people. Thus, a high degree of error variance indicates the potential for poor measurement. Using the standard formula for variance, we can also compute a variance for the observed scores (s_o^2) and a variance for the true scores (s_t^2), as shown in Table 5.1a.

Assuming that an individual's observed score is the sum of the individual's true score and error score (i.e., $X_o = X_t + X_e$), it follows that the total variance of the observed scores from a group of individuals will equal the sum of their true score and error score variances:

$$s_o^2 = s_t^2 + s_e^2. \tag{5.3}$$

If you examine the observed score variance (s_o^2) in Table 5.1a, you will see that it is indeed the sum of true score variance and error score variance:

$$s_o^2 = 291.67 + 316.67,$$
$$s_o^2 = 608.33.^*$$

You may have noticed that Equation 5.3 seems inconsistent with the formula for the variance of a composite variable that was introduced in Chapter 3. In Chapter 3, we showed that the variance for a set of variables is equal to the sum of variances *plus* a term that represents the extent to which the individual variables are correlated with each other. In fact, an observed score is a composite variable—it is the sum of two variables (i.e., a true score variable and an error score variable). Thus, you might expect that the variance of observed scores should be

$$s_o^2 = s_t^2 + s_e^2 + 2r_{te}s_t s_e. \tag{5.4}$$

In words, total observed score variance should be equal to true score variance plus error variance *plus* the covariance of true scores and error scores ($c_{te} = 2r_{te}s_t s_e$). However, as described above, we assume that error is independent of true scores, which implies that the correlation between error score and true scores is zero ($r_{te} = 0$). Therefore, the far-right term of the above expression, the covariance, will equal zero and will drop out of the equation, leaving us with

$$s_o^2 = s_t^2 + s_e^2 \ .$$

Equation 5.3 is a critically important formula in the classical theory of reliability. As we will discuss below, reliability will be defined in various ways in terms of the relationship between observed score, true score, and error score variance.

NOTE: *The discrepancy between 608.33 and 608.34 is due to rounding, see Table 5.1a.

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 89

Four Ways to Think of Reliability

In classical test theory, there are at least four ways to think about reliability. In one way or another, each of these conceptual approaches arises from the associations among observed scores, true scores, and measurement error, as described above. At one level, the approaches differ only with respect to the methods used to algebraically manipulate the terms associated with these variances. At another level, they represent different ways of conceptualizing or characterizing the concept of reliability.

Table 5.2 A 2 × 2 Framework for Conceptualizing Reliability

Statistical Basis of Reliability in Terms of . . . \ *Conceptual Basis of Reliability: Observed Scores in Relation to . . .*	*True Scores*	*Measurement Error*
Proportions of Variance	*Reliability is the ratio of true score variance to observed score variance* $R_{xx} = \frac{s_t^2}{s_o^2}$	*Reliability is the lack of error variance* $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
Correlations	*Reliability is the (squared) correlation between observed scores and true scores* $R_{xx} = r_{ot}^2$	*Reliability is the lack of correlation between observed scores and error scores* $R_{xx} = 1 - r_{oe}^2$

As shown in Table 5.2, these four approaches reflect two distinctions in the conceptualization of reliability. One distinction is whether an approach conceptualizes reliability in terms of "proportion of variance" or in terms of correlations. A second distinction is whether an approach conceptualizes reliability in terms of observed scores as related to true scores or to measurement error.

There are at least two reasons to become familiar with these different ways of thinking about reliability. First, an appreciation of the concepts expressed through the different approaches should help you develop a deeper understanding of the general meaning of reliability. Second, in your readings and discussions about tests and their reliabilities, you are likely to find that different people discuss reliability in different ways. Being familiar with these different perspectives and knowing how they are related to each other might help you avoid confusion when confronted with them in these discussions.

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 90

Reliability as the Ratio of True Score Variance to Observed Score Variance

Probably the most common expression of reliability is that it is the proportion of observed score variance that is attributable to true score variance:

$$R_{xx} = \frac{s_t^2}{s_o^2}, \tag{5.5}$$

where R_{xx} is the reliability coefficient. For example, for the responses presented in Table 5.1a:

$$R_{xx} = \frac{291.67}{608.33},$$

$$R_{xx} = .48.$$

This value tells us that about 48% of the differences that we see among respondents' observed scores can be attributed to differences among their true trait levels.

The size of the reliability coefficient indicates a test's reliability. Reliability ranges between 0 and 1, and larger R_{xx} values indicate greater psychometric quality. This is the case because, as R_{xx} increases, a greater proportion of the differences among observed scores can be attributed to differences among true scores. Notice that if true score variance is zero, then $R_{xx} = 0$. That is, an R_{xx} of zero means that everyone has the same true score. This underscores the fact that reliability is intrinsically tied to differences among people—if respondents do not differ in the characteristic being assessed by a test (i.e., if $s_t^2 = 0$), then the test's reliability is zero. In contrast, if true score variance is equal to observed score variance, then $R_{xx} = 1.0$. This would indicate that there is absolutely no measurement error affecting observed scores. In reality, measurement error always occurs to some degree.

Although there is no clear cutoff value separating good reliability from poor reliability, the reliability of .48 for the data in Table 5.1a is rather low. A perfect reliability ($R_{xx} = 1.0$) will not occur, but we would be much more satisfied with a reliability of .70 or .80 for research purposes. We would be worried if less than half of the variance in observed scores could be attributed to true scores.

Therefore, the test user who used the self-esteem questionnaire for the data in Table 5.1a might wish to improve the questionnaire's reliability. Imagine that she revised the questionnaire by rewriting some of its items—for example, by clarifying potentially ambiguous wording and by making sure to refer to the way that people "generally" feel about themselves. She hopes that such revisions will improve the reliability of the questionnaire. Furthermore, imagine that she asked the same six respondents to complete the revised version of the questionnaire. These hypothetical responses are presented in Table 5.1b. Did her revisions improve the psychometric quality of the self-esteem questionnaire?

Take a moment to contrast the data in Table 5.1a (the original questionnaire) and 5.1b (the revised questionnaire). First, notice that the individual's true scores

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 91

are the same for the revised test as they were in the original test. This occurs because the questionnaire is a measure of self-esteem, and we assume that individuals' true levels of self-esteem are stable across the two testing occasions. That is, self-esteem is a trait that is generally quite stable. Although people may experience temporary fluctuations in their self-esteem, we assume that each person has an overall level that reflects his or her typical level of self-regard. The self-esteem questionnaire is intended to measures these stable levels of self-esteem.

Second, notice the differences among the respondents. Again, let us focus on Ashley and Bob. As we have already pointed out for the original questionnaire data, there was a clear inconsistency between Ashley's and Bob's true score difference and their observed score difference. Specifically, Ashley's true score is 10 points higher than Bob's true score, but her observed score was 25 points *lower* than Bob's observed score. This reflects a substantial effect of measurement error. In contrast, their observed scores on the revised questionnaire seem to be much more consistent with their true scores. Specifically, Ashley's observed score is 5 points higher than Bob's observed score. Although this 5-point difference is still somewhat inconsistent with the full 10-point difference in their true scores, it is a relatively small inconsistency. Furthermore, the observed score difference on the revised questionnaire is at least consistent with the *direction* of the difference in their true scores. That is, the revised test produced scores in which Ashley scored higher than Bob, which is consistent with their true score difference. Thus, we begin to get a sense that the revised test does a better job of reflecting the true differences among respondents, as compared to the original test. This sense is confirmed when we compute the reliability for the revised test:

$$R_{xx} = \frac{291.67}{408.33},$$

$$R_{xx} = .71.$$

For the revised questionnaire, 71% of observed score variance can be attributed to variance in the true scores. The reliability of the revised questionnaire is much better than the reliability of the original questionnaire. This suggests that the item revisions paid off and that future test users should probably work with the revised test.

Lack of Error Variance

A second way of conceptualizing reliability is in terms of a lack of measurement error. We have already seen that error variance (s_e^2) represents the degree to which error affects different people in different ways—artificially inflating some people's scores and artificially deflating other people's scores. These effects obscure the true differences among people, as shown in our comparisons of Ashley and Bob. Therefore, reliability can be seen as the degree to which error variance is minimal in comparison to the variance of observed scores. We can state this formally, as we now demonstrate.

05(1)-Furr-45314-quark-4.qxd 8/30/2007 11:02 AM Page 92

In the previous section, we stated that reliability can be seen as the proportion of observed score variance that is attributable to true score variance:

$$R_{xx} = \frac{s_t^2}{s_o^2}. \tag{5.6}$$

We have also stated that observed score variance is the sum of true score variance and error variance (Equation 5.3):

$$s_o^2 = s_t^2 + s_e^2.$$

Rearranging terms algebraically, reveals that

$$s_t^2 = s_o^2 - s_e^2.$$

Substituting this into the numerator of Equation 5.6, we obtain

$$R_{xx} = \frac{s_o^2 - s_e^2}{s_o^2}.$$

Again rearranging:

$$R_{xx} = \frac{s_o^2}{s_o^2} - \frac{s_e^2}{s_o^2}.$$

And simplifying:

$$R_{xx} = 1 - \frac{s_e^2}{s_o^2}. \tag{5.7}$$

Note that

$$\left(\frac{s_e^2}{s_o^2}\right)$$

represents the proportion of observed score variance that is a function of error variance. Reliability is relatively high when this proportion is relatively small. That is, reliability is high when error variance is small in comparison to observed score variance.

For the data from the original self-esteem questionnaire:

$$R_{xx} = 1 - \frac{316.67}{608.33},$$

$$R_{xx} = 1 - .52,$$

$$R_{xx} = .48.$$

05(2)-Furr-45314.qxd 8/30/2007 5:37 PM Page 93

Thus, 52% of the variance in respondents' observed scores on the original questionnaire is produced by measurement error, leaving only 48% attributable to true score differences among the respondents.

What would a small degree of error variance indicate? It would indicate that respondents' scores are being affected only slightly by measurement error. More accurately, it would indicate that the error affecting one person's score is not very different from the error affecting another person's score. We see this in the data for the revised self-esteem questionnaire, where the error scores range only from –15 to +10. In addition, the standard deviation of error scores on the revised questionnaire is 10.80, indicating that the average person's error score is only about 11 points. In fact, measurement error accounts for only 29% of the variance in observed scores on the revised questionnaire. Contrast this with the data for the original self-esteem questionnaire, where error scores ranged from –15 to 25—a noticeably wider range of scores. In addition, the standard deviation of error scores on the original questionnaire is 17.80, indicating that the average person's error score is about 18 points. These facts reflect the greater effects of error in the original questionnaire, accounting for fully 52% of the variance in observed scores. Of course, if there is no error variance, then 100% of the observed variance in test scores will be associated with true score variance, and the test will be perfectly reliable.

The (Squared) Correlation Between Observed Scores and True Scores

This chapter began by stating that reliability is the degree to which differences in observed scores are consistent with differences in true scores. In Chapter 3, we saw that the correlation coefficient tells us the degree to which differences in one variable are consistent (i.e., correspond with) with differences in another variable. Thus, reliability can be seen in terms of the (squared) correlation between observed scores and true scores:

$$R_{xx} = r_{ot}^2. \tag{5.8}$$

Again, looking at the data in Table 5.1a, we have calculated the correlation between the observed scores and the true score, $r_{ot} = .69$. If we square this value, we get $r_{ot}^2 = .48$, which is R_{xx} as demonstrated earlier. The (unsquared) correlation between observed scores and true scores is sometimes called the "index of reliability" (Ghiselli et al., 1981). Please do not let this confuse you. If you square the index of reliability, you obtain the coefficient of reliability. When people refer to reliability, they typically are referring to the "coefficient" of reliability (R_{xx}). Only rarely will you hear people refer to the index of reliability (r_{ot}); however, an understanding of their connections should provide deeper insight into the concept of reliability.

We will take a moment to prove that the squared correlation between observed scores and true scores (r_{ot}^2) equals the ratio of true score variance and observed score variance

$$\left(\frac{s_t^2}{s_o^2}\right),$$

05(2)-Furr-45314.qxd 8/30/2007 5:37 PM Page 94

which is the most common way of conceptualizing the reliability coefficient. Recall from Chapter 3 that a correlation can be seen as a covariance divided by the product of two standard deviations:

$$r_{xy} = \frac{c_{xy}}{s_x s_y}.$$

Thus, the correlation between observed scores and true scores is

$$r_{ot} = \frac{c_{ot}}{s_o s_t}. \tag{5.9}$$

The covariance between observed scores and true scores is

$$c_{ot} = \frac{\sum(X_o - \overline{X}_o)(X_t - \overline{X}_t)}{N}. \tag{5.10}$$

From Equation 5.1, we assume that

$$X_o = X_t + X_e .$$

And because the mean error score is assumed to be zero (i.e., $\overline{X}_e = 0$, as explained above), the mean observed score is equal to the mean true score:

$$\overline{X}_o = \overline{X}_t .$$

Inserting this and Equation 5.1 into the covariance (Equation 5.10):

$$c_{ot} = \frac{\sum(X_t + X_e - \overline{X}_t)(X_t - \overline{X}_t)}{N}.$$

Algebraically simplifying this equation, we find that the covariance between observed scores and true scores is equal to the sum of (a) the variance in true scores and (b) the covariance between true scores and error scores:

$$c_{ot} = s_t^2 + c_{et} .$$

However, as we explained above, we also assume that error scores and true scores are independent, which means that they are not correlated with each other (i.e., $r_{te} = 0$ and therefore $c_{te} = 0$). So, the covariance between observed scores and true scores is simply equal to the variance in true scores:

$$c_{ot} = s_t^2. \tag{5.11}$$

05(2)-Furr-45314.qxd 8/30/2007 5:37 PM Page 95

Returning to the correlation between true scores and observed scores (Equation 5.9), we insert Equation 5.11 into the numerator:

$$r_{ot} = \frac{s_t^2}{s_o s_t}.$$

Simplifying this, we find

$$r_{ot} = \frac{s_t}{s_o}.$$

Squaring this, we find that that the squared correlation between observed scores and true scores is exactly equal to the ratio of true score variance and observed score variance:

$$r_{ot}^2 = \frac{s_t^2}{s_o^2}.$$

Thus, reliability can be seen as the squared correlation between observed scores and true scores. A reliability of 1.0 would indicate that differences among respondents' observed test scores are perfectly consistent with the differences among their true scores. A reliability of 0.0 would indicate that differences among respondents' observed test scores are totally inconsistent with the differences among their true scores. In such a case, the test is completely useless as a measure of a psychological characteristic. In practice, reliability is usually in between these two extremes.

Lack of (Squared) Correlation Between Observed Scores and Error Scores

Paralleling the previous ways of conceptualizing reliability, reliability can also be seen as the degree to which observed scores are uncorrelated with error scores. To the degree that differences in observed test scores reflect differences in the effects of error (instead of true scores), the test is unreliable. Thus,

$$R_{xx} = 1 - r_{oe}^2\,, \tag{5.12}$$

where r_{oe}^2 is the squared correlation between observed scores and error scores.

Once again, the data in Table 5.1a demonstrate this equivalence. We have calculated the correlation between observed scores and error scores (r_{oe} = .72). The square of this value is .52, which is equal to the ratio of error variance to observed score variance

$$\left(\frac{s_e^2}{s_o^2}\right)$$

05(2)-Furr-45314.qxd 8/30/2007 5:37 PM Page 96

As shown earlier, one minus this value is equal to the reliability:

$$
\begin{aligned}
R_{xx} &= 1 - r_{oe}^2 , \\
R_{xx} &= 1 - (.72)^2 , \\
R_{xx} &= 1 - .52 , \\
R_{xx} &= .48.
\end{aligned}
$$

We will algebraically prove that the squared correlation between observed scores and error scores (r_{oe}^2) equals the ratio of error score variance to observed score variance

$$\left(\frac{s_e^2}{s_o^2}\right)$$

The correlation between observed scores and error scores is

$$r_{oe} = \frac{c_{oe}}{s_o s_e}. \tag{5.13}$$

The covariance between observed scores and true scores is

$$c_{oe} = \frac{\sum(X_o - \overline{X}_o)(X_e - \overline{X}_e)}{N}. \tag{5.14}$$

Once again, from Equation 5.1, we assume that

$$X_o = X_t + X_e.$$

And because the mean error score is assumed to be zero (i.e., $\overline{X}_e = 0$), the mean observed score is equal to the mean true score:

$$\overline{X}_o = \overline{X}_t .$$

Inserting this and Equation 5.1 into the covariance (Equation 5.14):

$$c_{oe} = \frac{\sum(X_t + X_e - \overline{X}_t)(X_e - \overline{X}_e)}{N}.$$

Algebraically simplifying this equation, we find that the covariance between observed scores and error scores is equal to the variance of error scores:

$$c_{oe} = s_e^2. \tag{5.15}$$

Returning to the correlation between error scores and observed scores (Equation 5.13), we insert Equation 5.15 into the numerator:

05(3)-Furr-45314.qxd 8/30/2007 11:10 AM Page 97

$$r_{oe} = \frac{c_{oe}}{s_o s_e},$$

$$r_{oe} = \frac{s_e^2}{s_o s_e}.$$

Simplifying this, we find

$$r_{oe} = \frac{s_e}{s_o}.$$

Squaring this, we find that that the squared correlation between observed scores and error scores is exactly equal to the ratio of error score variance and observed score variance:

$$r_{oe}^2 = \frac{s_e^2}{s_o^2}.$$

Thus,

$$R_{xx} = 1 - r_{oe}^2 = 1 - \frac{s_e^2}{s_o^2}.$$

Perhaps the best way to think about this is to realize that if the correlation (r_{oe}) between observed scores and error scores is zero, then R_{xx} will equal 1.0. As the correlation of observed scores with error scores increases, the size of R_{xx} will decease. For example, compare the data in Table 5.1a (the original self-esteem questionnaire) to the data in Table 5.1b (the revised self-esteem questionnaire). For the original questionnaire, the correlation between observed scores and error scores was relatively large ($r_{oe} = .72$), resulting in a relatively low reliability ($R_{xx} = .48$). In contrast, the revised questionnaire produced responses with a relatively small correlation between observed scores and error scores ($r_{oe} = .53$), resulting in a higher reliability ($R_{xx} = .71$). Thus, reliability will be relatively strong when observed scores have relatively low correlations with error scores.

Reliability and the Standard Error of Measurement

The reliability coefficient is a useful number, particularly for comparing the reliabilities of several different psychological tests. For example, you might have two self-esteem tests, and you might want to know which is the more reliable. The reliability coefficient does not, however, directly address the problem of indexing the size of measurement error associated with a test. The R_{xx} can tell us which test is most reliable, but it does not tell us, in test score units, the average size of error scores that we can expect to find when a test is administered to a group of people.

05(3)-Furr-45314.qxd 8/30/2007 11:10 AM Page 98

As we will see later, the size of measurement error has important applications for interpreting the accuracy of test scores and for computing probabilities of scores in testing and research settings.

The standard deviation of error scores is a useful way of expressing the amount of error affecting responses to a test. Let us take a moment to think about the error score standard deviations for the two versions of the self-esteem questionnaire. For the original version (Table 5.1a), the error score standard deviation (s_e) is 17.80, which represents the average size of the absolute values of the error scores. In this case, 17.80 tells us that, on average, respondents' observed scores deviated from their true scores by nearly 18 points. However, if you examine the error standard deviation for the revised questionnaire (Table 5.1b), you will see that it is smaller than the error standard deviation for the original questionnaire. In this case, the error standard deviation is 10.80, indicating that respondents' observed scores on the revised questionnaire deviated from their true scores by only about 11 points. Thus, observed scores on the revised questionnaire are more accurate (i.e., closer to true scores) than observed scores on the original questionnaire.

The standard deviation of error scores has a special name; it is called the *standard error of measurement* (se_m), and it is one of the most important concepts in measurement theory. The standard error of measurement represents the average size of the error scores. The larger the standard error of measurement, the greater the average difference between observed scores and true scores and the less reliable the test.

As you might imagine, a test's standard error of measurement is closely linked to its reliability. In fact, as we will see later, we will need to estimate the se_m from an estimate of reliability. We can use reliability (R_{xx}) to find the standard error of measurement (se_m):

$$se_m = s_o\sqrt{1 - R_{xx}}, \tag{5.16}$$

where s_o is the standard deviation of the observed scores. Looking at the data from Table 5.1a, we see that the standard deviation of the observed scores is 24.66, and the reliability is .48. Thus, the se_m is

$$se_m = s_o\sqrt{1 - R_{xx}},$$

$$se_m = 24.66\sqrt{1 - .48},$$

$$se_m = 24.66(.72),$$

$$se_m = 17.80.$$

This value (17.80) is exactly equal to the standard deviation that is computed directly from the error scores (s_e). To prove that

$$se_m = s_o\sqrt{1 - R_{xx}},$$

remember that

$$R_{xx} = \frac{s_t^2}{s_o^2},$$

05(4)-Furr-45314.qxd 8/30/2007 11:11 AM Page 99

which is equivalent to (see Equation 5.7)

$$R_{xx} = 1 - \frac{s_e^2}{s_o^2},$$

which equals

$$\frac{s_e^2}{s_o^2} = 1 - R_{xx}.$$

Multiplying by s_o^2:

$$s_e^2 = s_o^2 (1 - R_{xx}).$$

Because $s_e^2 = se_m$ and taking the square root:

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

This shows how the standard error of measurement is related to R_{xx}. Notice that if $R_{xx} = 1$, then $se_m = 0$ and that se_m can never be larger than s_o. To reiterate, we will soon see that the standard error of measurement is an important psychometric value with implications for applied measurement.

Parallel Tests

If you have been paying close attention to this point in our discussion of reliability, you might be aware of an unpleasant fact. So far, reliability theory has been framed in terms of true scores, error scores, and observed scores. In contrast to the elegant theory of reliability, the reality of measurement is of course that we have no way of knowing people's true scores on a psychological variable or the error associated with their test responses. Thus, it may appear that there is no way to translate reliability theory into the actual practice of measurement. It may appear that we cannot actually evaluate a test's reliability or its standard error of measurement.

Classical test theorists avoid this problem by making one more assumption. *They assume that two psychological tests can be constructed in such a way that they are "parallel."* A pair of tests is considered *parallel* if all of the previous assumptions from CTT are true and if the following two additional assumptions hold true:

1. The tests measure the same psychological construct (this condition is known as "tau equivalence"). That is, participants' true scores for one test are exactly equal to their true scores on the other test.

2. The tests have the same level of error variance.

05(4)-Furr-45314.qxd 8/30/2007 11:11 AM Page 100

A consequence of these assumptions is that the observed scores on the tests will have the same mean, and the observed scores on the tests will have the same standard deviation. If two tests are parallel, then we will be able to compute a reliability coefficient and a standard error of measurement from their observed scores.

Imagine that you have two questionnaires that you think are measures of self-esteem—call them *X* and *Y*—and that you ask a group of people to take both tests. If the questionnaires both measure the same psychological construct (presumably self-esteem in this case), and if they have the same error variance (i.e., $s_{x_e} = s_{y_e}$), then *X* and *Y* are parallel tests. Notice that the hypothetical self-esteem questionnaires presented in Tables 5.1a and 5.1b are *not* parallel. Although their true scores are the same (i.e., they are measuring the same construct) and their observed means are identical, they have different error variances, which creates differences in the standard deviations of their observed scores. Thus, they fail to meet one of the assumptions for parallel tests.

However, if two tests—*X* and *Y*—are parallel, then we can compute the correlation in the usual way between the scores on the two tests. For example, if 100 people take tests *X* and *Y,* then each person will have two scores, and we can calculate the correlation between the observed test scores using the correlation coefficient described in Chapter 3 (r_{xy}). According to classical test theory, *the correlation between parallel tests is equal to reliability.* We can show that r_{xy} equals R_{xx} given the assumptions of classical test theory.

First, recall again that the correlation between the observed scores on the two tests is the covariance of the tests divided by the product of their standard deviations:

$$r_{xy} = \frac{c_{xy}}{s_x s_y}.$$

If the two tests are parallel, then by definition, their observed scores have equal standard deviations (i.e., if $s_{x_e} = s_{y_e}$, then $s_{x_o} = s_{y_o}$, which we simply call s_o:

$$r_{xy} = \frac{c_{xy}}{s_o s_o},$$
$$r_{xy} = \frac{c_{xy}}{s_o^2}.$$

Recalling that the definition of an observed score is $X_o = X_t + X_e$, then the observed scores can be seen as composite variables (each one being the composite of two components). Therefore, the covariance between the tests' observed scores is a covariance between composite variables. Furthermore, the covariance between two composite variables is the sum of the covariances between the components within the composite variables:

$$r_{xy} = \frac{c_{x_t y_t} + c_{x_t y_e} + c_{x_e y_t} + c_{x_e y_e}}{s_o^2},$$

05(4)-Furr-45314.qxd 8/30/2007 11:11 AM Page 101

where $c_{x_t y_t}$ is the covariance between true scores on test *X* and true scores on test *Y*, $c_{x_t y_e}$ is the covariance between true scores on test *X* and error scores on test *Y*, $c_{x_e y_t}$ is the covariance between error scores on test *X* and true scores on test *Y*, and $c_{x_e y_e}$ 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, errors scores are uncorrelated with true scores. In addition, error scores on test *X* are uncorrelated with error scores on test *Y*. Consequently, the three covariances that include error scores ($c_{x_t y_e}$, $c_{x_e y_t}$, and $c_{x_e y_e}$) are all equal to 0, so that

$$r_{xy} = \frac{c_{x_t y_t}}{s_o^2}.$$

Because true scores are equal for the two tests (i.e., each respondent's $X_t = Y_t$), the covariance between the true scores equals the variance of true scores (i.e., $c_{x_t y_t} = s_t^2$). This derives from the fact that a variance can be seen as the covariance between a variable and itself. Therefore, the correlation between parallel tests is

$$r_{xy} = \frac{s_t^2}{s_o^2}.$$

That is, the correlation between scores on parallel tests is equal to the ratio of true score variance to observed score variance, which is a definition of reliability (R_{xx}). The parallel test assumption will be crucial in the next chapter, when we discuss the procedures used to estimate the reliability in real-life testing situations.

Domain Sampling Theory

Domain sampling theory of reliability was developed in the 1950s as an alternative to classical test theory (Ghiselli et al., 1981). Domain sampling theory is an alternative to classical test theory in that both approaches arrive at the same place regarding procedures for calculating reliability, but they arrive there from different directions. For example, in classical test theory, reliability rests on the assumption that it would be possible to create two tests that are at a minimum parallel to each other. In domain sampling theory, you do not have to make this assumption, but if you follow the logic of the theory, you will end up with parallel tests by *fiat.*

Domain sampling theory rests on the assumption that items on any particular test represent a sample from a large indefinite number or domain of potential test items. Responses to each of these items are thought to be a function of the same psychological attribute. For example, suppose you had a test of self-esteem with 10 questions on the test. Differences in responses to each of these 10 questions should be related to differences in self-esteem among the people taking the test. Furthermore, the particular items on the test are thought to be a random sample from a population or domain of similar items, each of which is a measure of self-esteem.

05(4)-Furr-45314.qxd 8/30/2007 11:11 AM Page 102

If you created a test by selecting *N* items at random from a domain of items and then created a second test by selecting another set of *N* items at random from the same domain, in the long run, these pairs of tests should have the same mean and same standard deviation; in other words, on average, all test pairs selected in this fashion should be parallel to each other. If you have two parallel tests, then the test scores should correlate strongly with each other. The extent to which they do not correlate strongly with each other should be due to item sampling error. From this perspective, reliability is the average size of the correlations among all possible pairs of tests with *N* items selected from a domain of test items. The logic of domain sampling theory is the basis for a contemporary approach to reliability, called generalizability theory. We will explore this in greater detail in Chapter 12.

Summary

In this chapter, we have examined the theory of reliability from the perspective of classical test theory. Although there are other perspectives on reliability, classical test theory is the most well known, and it serves as the basis for many psychometric evaluations of psychological measures.

Classical test theory rests on a few fundamental assumptions about test scores and the factors that affect them. As we have described, classical test theory is based on the assumption that observed scores on a test are a simple additive function of true scores and measurement error (i.e., $X_o = X_t + X_e$). In addition, classical test theory rests on the assumption that measurement error occurs as if it is random. The randomness assumption has several implications—for example, error is uncorrelated with true scores, error averages to zero, and error on one test is uncorrelated with error on another test.

These assumptions have important implications for the nature of variability among test scores. As this book has emphasized, the meaning of psychological measurements is tied closely to the need to detect and quantify differences among people. Thus, variability among observed scores is the sum of true score variability and error variability. That is, the differences among people's observed scores arise from differences in their true scores and differences in the degree to which error affects their responses.

From this perspective, reliability reflects the links between observed scores, true scores, and error. As we have discussed, there are at least four ways to conceptualize reliability. Reliability can be seen in terms of variance. It is the ratio of true score variance to observed score variance, and it can also be seen as a lack of error variance. That is, reliability is high when the differences among participants' observed scores primarily reflect differences among their true scores. Reliability can also be seen in terms of consistency and correlations. It is the degree to which observed scores are correlated with true scores, and it can be seen as the degree to which observed scores are uncorrelated with error scores. That is, reliability is high when the differences among participants' observed scores are consistent with the differences among their true scores.

05(4)-Furr-45314.qxd 8/30/2007 11:11 AM Page 103

This chapter also discussed the standard error of measurement and the notion of parallel tests. These two concepts, which emerge from classical test theory, will be important tools when we translate the theory of reliability into the practice of psychometric evaluation of real test data.

Indeed, this chapter has focused on the theoretical basis of reliability. In order to illustrate the rather technical concepts, we have pretended to be omniscient, knowing respondents' true scores and the nature of error that affected each score. Of course, we will not know such things when we work with real test responses. Therefore, we can never really know the reliability of a test (just as we can never really know an individual's level of self-esteem). However, the notion of parallel tests will allow us to actually estimate test reliability with real data. The next chapter describes these estimation processes.

Suggested Readings

The classic in the development of classical test theory:

Gulliksen, H. (1950). *Theory of mental tests.* New York: John Wiley.

For a detailed discussion of domain sampling theory:

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: W. H. Freeman.