CHAPTER 7

Reliability of Test Scores and Test Items

In this chapter, the bane of test developers will be presented. This is the reliability of the test instrument. *Reliability* is an umbrella term under which different types of score stability are assessed. It is up to the test developer and producer to ensure that the appropriate reliability indices are reported. It is up to test consumers to know how to interpret the presented reliability information.

In essence, the reliability index of a test score indicates its stability. That may mean stability of test scores over time (test-retest), stability of item scores across items (internal consistency), or stability of ratings across judges, or raters, of a person, object, event, and so on (interrater reliability). The focus of this chapter will be on the stability of test scores and test items.

These approaches to reliability stem from models of classical test theory (CTT). One of the critical features of reliability from this perspective is that it is concerned solely with random measurement error. Before presenting the various reliability indices, recall that classical testing approaches assume that the raw score (X) on a test is made up of a true component (T) and a random error (E) component:

$$X = T + E. \tag{7–1}$$

The less random error, the more the raw score represents the true score. Because true scores are theoretical, a formula is needed that will allow for the calculation of the reliability without a true score component. We reviewed in Chapter 5 how a sample of raw scores can be used to get to the formula:

$$\text{Reliability} = 1 - [\text{VAR}(E)/\text{VAR}(X)]. \tag{7–2}$$

That is, the reliability is equal to the ratio of random error variance to total score variance.

In addition to the theory of true and error scores, we will also be adopting the domain-sampling model of CTT. This model assumes that the test is made up of a selection of a sample of parallel items from the universe of possible items. That is, there are assumed to be an infinite number of possible items that could be designed to assess any construct. A single test of 20 items, for example, is a sample from that universe.

Different ways to assess reliability using CTTs will be presented. Examples of how to calculate the various types of reliability indices and computer program printouts and their interpretations will be provided.

Test-Retest Reliability

This index of reliability assesses the stability over time for a set of scores on a particular test for a given sample. This means that the same test is given to the same sample at a point in time (T1) and then again at a later point in time (T2). The approach has some obvious advantages. First, only one test is needed, reducing the cost of test item development. In addition, the items (stimuli) are the same from T1 to T2. This ensures that the same construct is measured exactly the same way both times.

The test-retest reliability index is simply the zero-order correlation between the test scores at T1 and T2. If one were to obtain exactly the same scores from the entire sample of respondents at both times, then the correlation would be perfect (1.0). Inevitably, measurement error comes into play, and scores will vary from T1 to T2. This might be due to random error, such as some participants having been feeling poorly on the day of the test at T1 and feeling well at T2 or the room having been overly warm at T2 compared to T1. Table 7.1 shows 20 scores obtained from a sample of employees on the 10-item Team Player Inventory (TPI) at T1 and T2. The resulting correlation between these two vectors of scores is 0.91. This indicates that the scores are quite stable from T1 to T2.

When interpreting any reliability index, it is important to note that the generated value is a *squared* value. That is, it is usually the case that if two variables are correlated with one another (say at 0.40), then the square of that value is equal to the shared variance between the two variables (i.e., $0.40^2 = 0.16$; the variables share 16% of their variance), but because of the manner in which the reliability index is derived, the zero-order correlation is already a squared value. Thus, it is concluded that 91% of the variance of the scores at T1 on the TPI is shared with the variance at T2. The output for the analysis is shown in Box 7.1.

The fluctuation that occurs in scores from T1 to T2 may be due to systematic change rather than random fluctuation. A systematic change would occur in the following example. Assume that at T1, a sample of employees is given an honesty test. These employees are now sensitized to the types of questions asked on the test (such as, Do you ever take office supplies home for personal use?). Then, when the test is given six weeks later at T2, some people's scores might be changed from T1 not because they changed in honesty but because they were sensitized to the honesty issue at their place of work. This phenomenon is called reactivity. Another

Table 7.1 Data for Calculating the Test-Retest Reliability of the Team Player Inventory for a Group of 20 Employees

Case Number	*Time 1 Scores*[a]	*Time 2 Scores*
1	40	42
2	30	41
3	20	26
4	17	18
5	45	39
6	27	23
7	15	22
8	32	28
9	33	33
10	45	41
11	41	37
12	30	32
13	25	22
14	12	11
15	49	43
16	48	47
17	34	30
18	36	32
19	26	29
20	22	20

a. Scores can range from 10–50.

example of this might occur in the following scenario. Assume a test of reading skill is given at T1 to a group of students. Then the students take a four-week reading skills course. They take the test again at T2 after the course is completed. Most of the scores would be expected to be different from T1 (in this case, increase, if the skills course does its work). This systematic change in scores is not considered to be error in CTT because it is not random.

So, one practical question that is often asked is, How long should one wait between test administrations to gather the T2 data? Because the exact same test is administered at T1 and T2, the potential for practice or carryover effects is substantial and this would artificially inflate the reliability index. Individuals not

Box 7.1 Test-Retest Correlation SPSS Output

		Time 1	*Time 2*
Time 1	Pearson Correlation	1	0.910
	Sig. (two-tailed)	—	0.000
	N	20	20
Time 2	Pearson Correlation	0.910	1
	Sig. (two-tailed)	0.000	—
	N	20	20

Note: Correlation is significant at the 0.01 level (two-tailed).

The printout shows that the correlation between Time 1 and Time 2 is 0.910 and is significant (0.000), and the sample size (*N*) is 20.

used to taking tests may do better at T2 simply because they have had an opportunity to practice taking the test at T1. In addition, there may be specific questions on the test that a respondent did not know the answer to at T1 (e.g., What is the meaning of *sanctimonious?*). The person may have looked up the answer after taking the test and therefore know the correct response at T2. Respondents may also remember what they gave as a response at T1 to particular items (e.g., Rate how much you like chocolate ice cream) and respond the same way at T2. All of these examples are problems that face test-retest reliability estimates.

Ghiselli, Campbell, and Zedek (1981) say, "It is desirable to maximize the interval between testing occasions to minimize the effects of memory" (p. 249). On the other hand, if the test assesses a construct that may be affected by historical/situational events (e.g., feelings toward a political party, levels of anxiety) or maturation/learning (e.g., cognitive ability) test-retest intervals that are overly long will likely produce reliability indices that are lower than would have been the case had the interval been shorter. This raises the issue of whether or not the test-retest reliability index should even be used in instances where the construct is susceptible to change.

So it seems that the test-retest method is most appropriate for tests that assess traits, such as intelligence or personality, that are assumed to be stable over time. In such instances, the maximal interval possible without undue cost is the rule of thumb to follow for the time interval between test administrations.

Clearly, there are some problems with the test-retest methodology. One is simply that there will likely be attrition in the sample—that is, there will be fewer individuals taking the test at T2 than T1. Participants will drop out, not show up, expire, be sick on the day of the second testing session, and so forth. Second, it is expensive to administer a test two times. Administration and scoring costs for some tests run into the hundreds of dollars. Third, the reactivity or sensitization discussed earlier can potentially negatively affect the reliability index. Fourth, the time interval may be inappropriate, unduly inflating or attenuating the reliability index.

Alternative Forms Reliability

To overcome, primarily, the problem of carryover effects and situational changes to test takers, the alternative forms approach to assessing reliability was developed. This is a much more costly approach in that two versions of the test must be developed. Extreme care must be taken to make sure that items are actually parallel, or equal, across test versions.

Specifically, alternative forms reliability data is collected first by one form of a test (Form A) when it is given to a sample at a point in time, T1. Then, at a later point in time (about two weeks), T2, the alternative form (Form B) is given to the same participants. The zero-order correlation between the test scores on Form A and Form B provides the index of alternative forms reliability.

The question of equality of items is of primary concern with alternative forms reliability. In the past, expert opinion and comparisons of pass rates were the only ways to really demonstrate that two forms of the test were equivalent. However, IRT produces item parameters that much more definitively answer the question of item and test equality. The onus is on the test developer to ensure that the alternative forms are equivalent and, thus, the reliability index associated with them is reasonable.

Another concern, although mitigated somewhat by the shorter time period between testing sessions, is attrition; the sample is likely to be smaller at T2 than at T1 for a variety of reasons. And, as with test-retest reliability, the test needs to be administered on two separate occasions, which may be very costly, depending on the specific test in question.

Interpretation of the index follows similarly from that of test-retest reliability. In this instance, however, the stability of the scores is from one test form to another test form.

Measures of Internal Consistency

There are several measures of internal consistency. All of them assess the stability, or consistency, of responses across items and are primarily based on the intercorrelations between items. The stability indices are not across total test scores but across items. Tests that are speeded, or where the items are ordered in terms of difficulty, should not be subjected to internal consistency assessments of reliability as the correlations among items are directly affected by time constraints and difficulty, rendering them spuriously high (Nunnally & Bernstein, 1994).

One of the most useful features of measures of internal consistency is that they can be calculated based on a single sample with just one test administration. This is a very desirable feature and has encouraged their use.

Split-Half. The split-half correlation between halves of tests was the first measure of internal consistency. That is, the test was divided into two equal parts, the scores on the two halves were calculated, and then the correlation between the two halves provided the split-half reliability. One problem with the split-half method has been, Where should the test be split? Usually, items are randomly split or all even numbers and all odd numbers make up the two halves. It is not wise to use the first and second halves, as the test taker may be more nervous on the first half or fatigued on the second half.

An example of a data set of 15 people taking the TPI is shown in Table 7.2. Recall that there are 10 items, so we decide that Half 1 will be the total of items 1, 3, 5, 7, and 9, whereas Half 2 will be the total of items 2, 4, 6, 8, and 10. The resulting split-half reliability for this data set is 0.825. The output for this analysis is shown in Box 7.2.

One problem spotted early on with the split-half method was that it underestimated the actual reliability index. This is because, under CTT, tests that are longer (i.e., have more items) are more reliable (assuming items that are similar are added). Note that the split-half reliability index was calculated on just five items for each half in our example. The length of the whole test, which respondents did complete, was actually 10 items. A formula to estimate the reliability of a test that is longer than the one on which the split-half coefficient was generated is called the

Table 7.2 Data for Calculating the Split-Half Reliability of the Team Player Inventory for a Group of 15 Employees

Case Number	*Half 1 Scores*[a]	*Half 2 Scores**
1	12	12
2	16	16
3	6	8
4	9	13
5	18	15
6	13	14
7	12	17
8	11	14
9	10	10
10	23	20
11	11	14
12	7	10
13	12	10
14	19	15
15	20	17

a. Scores range from 5–25.

Box 7.2 Split-Half Correlation and Reliability SPSS Output

		Half 1	*Half 2*
Half 1	Pearson Correlation	1	0.825
	Sig. (two-tailed)	—	0.000
	N	15	15
Half 2	Pearson Correlation	0.825	1
	Sig. (two-tailed)	0.000	—
	N	15	15

Note: Correlation is significant at the 0.01 level (two-tailed).

The printout shows that the correlation between Half 1 and Half 2 is 0.825 and is significant (0.000), and the sample size (*N*) is 15.

If you analyze the data in the Reliability program of SPSS, you can indicate that you want to use the split-half method. In this case, the printout would indicate the following:

Reliability Coefficients

N of Cases = 15.0; *N* of Items = 2

Correlation Between Forms = 0.8249

Equal-Length Spearman-Brown = 0.9040

Note that the equal-length Spearman-Brown value of 0.9040 from the output is equal to the Spearman-Brown calculation done by hand.

Spearman-Brown prophesy formula. Whereas this is a general formula that can be used to assess a variety of different questions about test length and reliability, it is presented here because it is used extensively in calculating the "corrected" split-half reliability. The formula is

$$r_{cc'} = kr_{xx}/[1 + (k - 1)(r_{xx})], \tag{7–3}$$

where $r_{cc'}$ = the expected reliability, k = the proportion the test is changed (e.g., 2 if it is doubled in length, 0.5 if it was halved), and r_{xx} = the original reliability index.

So, in our example, the original reliability index was 0.825. The expected reliability if the number of items on the test was doubled would be

$$\begin{aligned} r_{cc'} &= 2(0.825)/[1 + (2 - 1)(0.825)], \\ &= 1.65/1.825, \\ &= 0.904. \end{aligned}$$

Thus, our corrected split-half reliability for this set of data is 0.904.

The general formula can be used in some instances to shorten a test by a fixed amount and estimate the loss in reliability. For example, suppose a test is 100 items in length and has a reliability index of 0.90. However, it is too long for test takers to complete without complaining of fatigue. If the test is shortened to 75 items, there will be a loss in the reliability. The amount of that loss can be calculated using the following:

$$\begin{aligned} r_{cc'} &= 0.75(0.90)/[1 + (0.75 - 1)(0.90)] \\ &= 0.675/0.775, \\ &= 0.871. \end{aligned}$$

Thus, if 25 items are dropped from the test (i.e., 0.75 is used for the proportion the test is shortened), the reliability is expected to drop from 0.90 to 0.87. It must be decided if the fatigue experienced by the test takers is worth giving up the higher reliability.

One other useful way to use the formula is to rearrange the terms to estimate how much a test must be lengthened to achieve a desired level of reliability. This formula is

$$k = [(r_{cc})(1 - r_{xx})]/[(r_{xx})(1 - r_{cc})]. \tag{7–4}$$

For example, assume a 50-item test has a reliability of 0.70 and increasing the reliability to 0.80 is desired. Using the formula, the proportion increase needed can be calculated:

$$\begin{aligned} k &= [(0.80)(1 - 0.70)]/[(0.70)(1 - 0.80)], \\ &= 0.24/0.14, \\ &= 1.71. \end{aligned}$$

The test would need to increase by 1.71 times. This means the test would have to be 86 items in length to achieve the desired reliability (i.e., 50 × 1.71 = 85.7).

Cronbach's alpha (α). The α coefficient (Cronbach, 1951) is probably the most pervasive of the internal consistency indices. In fact, it is so pervasive that it has almost become synonymous with reliability. Its correct classification, however, is as a measure of internal consistency. The formula for α is

$$\alpha = (N)(r_{\text{mean}})/[1+ (r_{\text{mean}})(N - 1)], \tag{7–5}$$

where N = the number of test items and r_{mean} = the average intercorrelations among the items.

Assume that there is a 20-item test where the average of the correlations of each item with all other items is 0.30. To find the α for this test, the formula would be

$$\begin{aligned} \alpha &= (20)(0.30)/[1 + (0.30)(20 - 1)], \\ &= 6/6.7, \\ &= 0.896. \end{aligned}$$

Thus, the α for the test is 0.896. Novick and Lewis (1967) demonstrated that α is the average of all possible combinations of split-half reliabilities. They also showed that, in general, α is the lower bound for a test of parallel items; that is, it is a conservative measure of internal consistency.

The equation shows that the value of α is dependent not only on the intercorrelations among the items but also on the length of the test. Note that if the test on which α was calculated was 10 items in length and not 20, the resulting α would be

$$\begin{aligned}&(10)(0.30)/[1 + (0.30)(10 - 1)],\\&= 3/3.7,\\&= 0.810.\end{aligned}$$

This is much lower than the original value despite the fact that the intercorrelations remained constant.

An example output for Cronbach's α is shown in Box 7.3. The data for this are 72 responses by employees to an Innovation Market Strategy scale that assesses employee perceptions about the market strategy being pursued by their organization (Kline, 2003). Interpretation of the results is also provided.

Coefficient Theta (θ). One little-used index of internal consistency is that of coefficient theta (θ; not to be confused with the theta of IRT). Theta is based on a principal components analysis of the test items. It differentially weights items that correlate more with each other than does θ. As a result, it is a special case of θ where a "weighting vector has been chosen so as to make alpha a maximum" (Carmines & Zeller, 1979, p. 61).

The formula for coefficient θ is

$$\text{coefficient } \theta = [(N/(N-1)]\ [1-(1/\lambda)], \tag{7–6}$$

where N = the number of items in the test and λ = the eigenvalue of the first principal component. For the moment, exactly where the eigenvalue comes from will be set aside as it will be discussed extensively in Chapter 10. At this point, just note that it is a measure of the variance shared between the items and it will be given in the examples in this chapter.

Assume that the responses by a sample of participants to the 10-item TPI scale is subjected to a principal components analysis. The eigenvalue for the first principal component is 6.7.

Coefficient θ for this data set can thus be calculated as follows:

$$\begin{aligned}\theta &= [20/(20 - 1)][(1 - (1/6.7)],\\&= 1.053(0.851),\\&= 0.896.\end{aligned}$$

The internal consistency for this set of items based on theta is 0.896. Coefficient θ is less conservative than α and is also less likely to be low if a test has a small number of items. It provides an alternative to α that can be readily computed.

Box 7.3 Cronbach's α SPSS Output

Item	*Scale Mean if Item Deleted*	*Scale Variance if Item Deleted*	*Corrected Item-Total Correlation*	*Alpha if Item Deleted*
M1	20.8750	26.1673	0.6350	0.8524
M4	22.2639	25.6336	0.6072	0.8545
M7	22.3611	26.1213	0.5756	0.8578
M10	21.1289	25.1354	0.7334	0.8422
M13	21.3889	23.8748	0.7396	0.8392
M16	21.5972	23.9059	0.6886	0.8452
M19	21.4306	25.0374	0.6293	0.8521
M22	21.6111	26.6072	0.4162	0.8770

Reliability Coefficients

N of Cases = 72.0; *N* of Items = 8

Alpha = 0.8689

Note that there are eight items and 72 cases used in this analysis. This scale has an internal consistency of 0.87, which is quite high. Examining the last column, notice that the α would drop if any of the items except the last one were to be deleted from the scale. That is, if M1 were deleted, then the α would drop from 0.8689 to 0.8524. This means that M1 contributes to making the internal consistency of the scale high. Note, however, that if the last item, M22, were to be deleted, then the α level would increase to 0.8773 (just slightly higher than the original 0.8689). Although this increase is not much and may very well be due to sampling error, it is useful to examine this column to see if any of the items can be deleted that would substantially increase the α level.

An example of the output for a principal components analysis that is used in calculating coefficient θ is shown in Box 7.4. The same data were used as that in demonstrating the Cronbach's α output. Interpretation of the results is provided.

Kuder-Richardson 21 (KR21). Kuder and Richardson introduced a measure of internal consistency for dichotomous items in 1937. They provided several formulae, the most notable one for our purposes being the KR21. It is as follows:

$$\text{(7–7)} \qquad \text{KR21 internal consistency} = [N/(N-1)]\ [1 - (\Sigma p_i q_i/\sigma^2)],$$

where N = the number of items, $p_i q_i$ = the proportion of individuals who pass (p) multiplied by the proportion of people who fail (q) each item (and then summed across all of the items), and σ^2 = the variance of the total test.

Box 7.4 Coefficient Theta Analysis: Total Variance Explained

First, a principal components analysis is carried out and the following SPSS output is generated.

Extraction Method: Principal Component Analysis

Component	*Initial Eigenvalues*	*% of Variance*
1	4.274	53.423
2	0.931	11.642
3	0.775	9.686
4	0.588	7.346
5	0.477	5.964
6	0.392	4.897
7	0.300	3.748
8	0.264	3.294

Note that the first eigenvalue is equal to 4.274. With eight items, the theta coefficient would be equal to (8/7)(1 – 1/4.274) = (1.143)(1 – 0.234) = (1.143)(0.766) = 0.876. This internal consistency value of 0.876 is just slightly higher than the one calculated using Cronbach's α.

There are several other parts to this output that we will be going over in the next chapter on validity.

So, for example, if 40 workers respond to a 10-item honesty test where individuals respond to statements as true (pass) or false (fail), the KR21 index can be calculated. Assume the variance of the test as a whole is 0.974 and the item pass and fail values are set up as in Table 7.3. The $\Sigma p_i q_i = 1.326$. So, substituting,

$$
\begin{aligned}
\text{KR21} &= [10/9]\ [1 - (1.326/0.974)], \\
&= (1.111)(0.3614), \\
&= 0.40.
\end{aligned}
$$

Thus, it would be concluded that the internal consistency of this 10-item scale is very low. The KR21 is provided as an option in most computer programs. Interpretation of the magnitude of the KR21 is similar to other measures of internal consistency.

Setting Confidence Intervals

Once the reliability coefficient is calculated (by whatever means), it can be used to set confidence intervals around a given score. This is particularly important in

Table 7.3 Item *p* and *q* Values for Calculating the Kuder-Richardson 21

Item	*p*	*q*	*pq*
1	0.175	0.825	0.1444
2	0.200	0.800	0.1600
3	0.125	0.875	0.1094
4	0.825	0.175	0.1444
5	0.925	0.075	0.0694
6	0.175	0.825	0.1444
7	0.175	0.825	0.1444
8	0.100	0.900	0.0900
9	0.125	0.875	0.1094
10	0.700	0.300	0.2100
			$\Sigma pq = 1.326$

applied settings where a decision-maker should have a particular level of confidence in the accuracy of the score that the respondent obtained and report that value. This process is called setting a confidence interval around a score. To set the confidence interval, the raw score of an individual (X), the reliability of the test (R), and the standard deviation of the test (SD) all must be known.

While most confidence interval setting procedures use the raw score around which the interval is set, Nunnally and Bernstein (1994) argue that the interval should be set about the individual's estimated true score (T'), which is determined by

$$T' = (R)(X). \tag{7–8}$$

The more reliable the test, the less likely the individual's estimated true score is to regress toward the mean of the distribution. Suppose a score of 120 on an IQ test is obtained, and the test has a reliability of 0.95 and a standard deviation of 12. The T', then, would be $(0.95)(120) = 114$.

To set the confidence interval around the score of 114, the standard error of measurement (SE) is required. It is calculated by

(7–9) $SE = SD\sqrt{(1 - \mathrm{R})}$. In this case, the SE would equal

$$SE = 12\sqrt{(1 - 0.95)},$$
$$= 2.68.$$

Then, if I want to be 90% confident that the score will fall within a certain range, the z value associated with the 90% confidence interval (1.65) will be multiplied with the value of the *SE* (1.65 × 2.68 = 4.42). Then this value is added to and subtracted from the estimated true score (i.e., 114 ± 4.42). This leads to the following two equations:

$$114 - 4.42 = 109.58 \text{ and}$$

$$114 + 4.42 = 118.42.$$

Thus, I can be 90% confident that if the test was administered to the test taker 100 times, 90 times out of 100, the true score would fall between 109.58 and 118.42.

In this example, the 90% confidence interval was set. If the 95% or 99% confidence intervals were desired, the *z* values to use would have been 1.96 or 2.58, respectively.

Reliability of a Composite

Sometimes it is of interest for test users to want to create a composite score from a variety of tests. For example, scores on four math subtests (components), addition, subtraction, multiplication, and division, could be added together to obtain a composite score of basic math fact mastery. The process for determining the reliability of the composite follows the same logic as does adding test items to a single test. That is, the more the components (whether they be items or whole tests) are correlated with one another, the higher the reliability of the composite when these components are added. The general formula for the reliability of a composite measure is

$$r_{\text{comp}} = 1 - [k - (kr_{ii\text{mean}})]/[k + (k^2 - k)\ r_{ij\text{mean}}], \tag{7–10}$$

where r_{comp} = the reliability of the composite, k = number of components, $r_{ii\text{mean}}$ = mean reliability of the components, and $r_{ij\text{mean}}$ = mean correlation between components.

Now, assume that the mean average reliability for the four component tests of addition, subtraction, multiplication, and division is 0.80, and the mean average correlation among the tests is 0.60. Substituting into the equation,

$$\begin{aligned} r_{\text{comp}} &= 1 - [4 - (4 \times 0.80)]/[4 + (16 - 4)0.60], \\ &= 1 - (0.80/11.2), \\ &= 0.93. \end{aligned}$$

Thus, the composite has a reliability that is higher than the average of the individual components (0.80). If the components were not highly intercorrelated, then the composite reliability would not have been much improved. For example, assume the mean correlation among the tests was 0.10 rather than 0.60. In this case, the resulting reliability of the composite would be 0.85.

This brings up the issue of whether or not to create a composite score from components that are heterogeneous in nature versus homogeneous. There is certainly a case to be made that if the components are homogeneous (have reasonably high intercorrelations), then creating a composite will provide a more reliable score and the meaning of the composite remains easily interpretable. The same is not true when combining heterogeneous scales. First, the reliability of the composite may be lower than each component alone, and second, interpreting a composite made up of a set of unrelated variables is "complicated if not impossible" (Allen & Yen, 1979, p. 224).

General intelligence tests often combine scores into a composite based on diverse constructs such as verbal fluency, pattern recognition, and spatial skills. Whether or not the resulting total score is really representative of "general intelligence" has been questioned. One positive aspect of these types of composite scores is that they tend to be related to other criterion variables that are also multidimensional (such as job or scholastic performance). In the end, make sure that the purpose of creating a composite score and the implications of doing so are known before going ahead to create such a variable.

Difference Scores—A Reliability Concern

A common event in research and practice with scores on items, or tests, is to calculate difference or change scores. For example, a training program for managers may be designed so that they learn to be good team leaders. To assess the effectiveness of the program, a team leadership scale is administered to them before training and then after training. Then the differences between the scores are calculated. These may then be used to relate to other variables of interest (such as subordinate satisfaction), presuming that those managers that changed the most should have better ratings of satisfaction by their subordinates. Another example is that employees might be asked to rate the actual culture of their organization on a supportiveness inventory and then asked to rate the desired supportiveness. The difference between them, then, is calculated and used.

A problem with this occurs in the CTT paradigm in that the reliability of difference scores is almost inevitably lower than the reliability of either of the component measures. The formula for the reliability of difference scores is as follows:

(7–11) $$r_{\text{diff}} = \{[(r_{xx} + r_{yy})/2] - r_{xy}\}/(1 - r_{xy}),$$

where r_{diff} = the reliability of the difference score, r_{xx} = the reliability of the first test, r_{yy} = the reliability of the second test, and r_{xy} = the correlation between the two tests.

Assume that the test of actual supportiveness from the example above had a reliability of 0.80 and the test of desired supportiveness had a reliability of 0.90. Further assume that the correlation between them was 0.50. Substituting in the formula,

$$\begin{aligned} r_{\text{diff}} &= \{[(0.80 + 0.90)/2] - 0.50\}/(1 - 0.50), \\ &= 0.35/0.50, \\ &= 0.70. \end{aligned}$$

The difference score reliability is 0.70—much lower than the reliability of either of the two component tests. This is because the two tests are moderately correlated. If the correlation between them was lower (e.g., 0.20 rather than 0.50), then the reliability of the difference score would be 0.81. If they were not correlated at all, then the resulting reliability would simply be the average of the two component reliabilities, in this case, 0.85.

This problem of difference score reliability occurs because of assumptions made in CTT. Theoretically, if two measured variables, *X* and *Y*, are highly correlated, then the "true scores" of *X* and "true scores" of *Y* are also assumed to be overlapping. If *X* and *Y* are measures of actual and ideal supportiveness, they will likely be correlated—that is, their true scores will overlap. What is not overlapping in the true scores is random error.

Assume the difference scores between actual and desired organizational supportiveness (the ones with the 0.70 reliability) are to be used by correlating them with employee intentions to leave the organization. The relationship between the difference scores and intentions to leave the organization will be attenuated due to the unreliability of the difference score. If a statistically significant relationship between the variables of interest is not found, it might be attributable to the unreliability of the difference score measure. It is worth noting that some researchers have provided explanations suggesting that the concern over the unreliability of difference scores is not as problematic as once thought (e.g., Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983; Williams & Zimmerman, 1996a, 1996b).

It is also important to examine the distributions of the scores on the two components. For example, assume one group of managers is sent into team training Program A and another group into team training Program B. Pretest and posttest measures on how much they learned from the training are obtained and the difference scores for each person are generated. Then the utility of Program A versus Program B is assessed by doing a *t* test on the difference scores of the two. It is assumed that the one with the greatest change is better. However, what if the group of managers that went into Program A all had relatively high scores on the pretest? If this was the case, then there was very little room for change for this group in advance of the training. Thus, Program A is disadvantaged for change.

It has been suggested in cases such as these that rather than difference scores, an analysis of covariance be used. In this instance, the covariate would be the pretest score, the dependent variable would be the posttest score, and the independent variable would be the training program group (Arvey & Cole, 1989).

Cribbie and Jamieson (2000) offer a compelling case for using structural equation modeling to assess difference score relationships with other variables of interest. The take-home message is simple: be wary of using difference scores as they can lead to potential problems in analysis and interpretation of findings.

Regardless of your own sentiments about difference scores, it is important to make cautious use and interpretation of results where they are used. If the two components are not strongly correlated and justification can be made for the difference score being a psychologically meaningful construct, then using difference scores might be warranted. It is also possible to correct the correlation between the difference score and the variable of interest using the correction for attenuation due to the unreliability of the difference score (Allen & Yen, 1979).

Practical Questions

There are a couple of practical questions that commonly occur regarding calculating reliability coefficients. One is, How many people should be in the sample? There really is not an answer to this question because the reliability index is descriptive, not inferential. Only in inferential statistical tests are issues of sample size important to test a statistic for significance. Rather, the question should be phrased, Is the sample on which the reliability coefficient is calculated representative of the samples on which I will want to use the test? Some test publishers provide reliability estimates for various samples (e.g., by gender, race, age), and this is very useful information as a consumer. As a test developer, you will want to assess reliability on samples of relevant populations. It makes no sense to develop a test of cognitive impairment for preschool children and test its reliability on elementary school children. So, sample size is less a concern than sample characteristics.

Another common question is, How high should the reliability index be to be considered "good"? This question is different for different test uses as well as for different types of reliability indices. Tests that will be used for making decisions about individuals (e.g., employee or student selection tests, tests that identify special needs students, tests that are used to determine treatment protocols) are expected to have very high reliability (0.90 at a bare minimum, as suggested by Nunnally and Bernstein, 1994). When individuals' lives are at stake, it is good practice not to accept any measurement error. One the other hand, in the early stages of research or where the test results will not be used for making decisions about people that affect their lives (sometimes called *low stakes*), modest reliability (0.70) of tests is acceptable (Nunnally & Bernstein, 1994). Another issue in reliability magnitude is the type of index used. For example, indices generated via the test-retest method are typically lower than are internal consistency methods for homogeneous sets of items.

Another question commonly presented is, What type of reliability analysis should I conduct? If the test contains homogeneous items, at the very least an internal consistency analysis needs to be conducted, unless it is a speeded test. In the latter case, alternative forms should be used. Test-retest is useful if the construct is not susceptible to historical, maturational, learning, or situational effects. Because of the potential problems with carryover, a long time should pass between test taking sessions. If the option for generating an alternative forms reliability index presents itself, then this would be a better way to assess reliability over time.

Each of the reliability estimate options has pros and cons associated with it. The test developer and test consumer should be testwise enough to know what type of reliability should be conducted and presented.

Summary and Next Steps

This chapter has focused on what is the primary concern for many psychometricians—reliability of the test scores. The topics that were covered included the following:

a. methods to assess reliability: test-retest, alternative forms, internal consistency, and composite scores;

b. applications of the reliability index in setting confidence intervals;

c. the potential problem of reliability with regard to difference scores; and

d. practical issues such as interpretation of indices, test length, sample sizes, and reliability standards.

In the next chapter, the discussion of reliability is extended to rater consistency. In addition, a note about modern test theory and how it differs from classical test theory in its approach to test reliability will be covered, as well as the issue of reliability generalization.

Problems and Exercises

1. Describe an example of test-retest reliability.
2. If the correlation between test scores at Time 1 and Time 2 is 0.85, how would this be interpreted?
3. What are some problems associated with reliability assessed via the test-retest method?
4. Under what circumstances is reliability assessed via the test-retest method most appropriate?
5. What are the strengths and drawbacks of alternative forms reliability?
6. Why is internal consistency such an easy way to assess reliability from a methodological perspective?
7. If a split-half reliability on a test is calculated to be 0.70, what would be the corrected split-half?
8. If a reliability of 0.60 is obtained on a test that is 10 items in length and the developer wants to increase it to 0.80, how many more items would need to be added to it?
9. If the average intercorrelation among all of the items on an eight-item test is 0.50, what is the Cronbach's alpha (α)?
10. If a principal components analysis on a set of six items is run and the first eigenvalue is equal to 4.0, what is the coefficient theta (θ)?
11. What would the KR21 be for a set of 20 dichotomous items where the overall test variance is 0.80 and the $\Sigma p_i q_i = 1.5$?
12. Set the 95% confidence interval around the true score of someone who obtained a raw score of 50 on a test that has a reliability of 0.90 and a standard deviation of 10.

13. What is the composite reliability of two tests that have a mean reliability of 0.70 and where the correlation between them is 0.50?

14. What would be the reliability of the difference score between two tests where the tests both had reliabilities of 0.80 and the correlation between the tests was 0.30?

15. If you obtained a reliability estimate of 0.80 on a test, how would you interpret it and use the test?