Module 5

Classical True Score Theory and Reliability

Any phenomenon we decide to "measure" in psychology, whether it is a physical or mental characteristic, will inevitably contain some error. For example, you can step on the same scale three consecutive times to weigh yourself and get three slightly different readings. To deal with this, you might take the average of the three weight measures as the best guess of your current weight. In most field settings, however, we do not have the luxury of administering our measurement instrument multiple times. We get one shot at it and we had better obtain the most accurate estimate possible with that one administration. Therefore, if we have at least some measurement error estimating a physical characteristic such as weight, a construct that everyone pretty much agrees on, imagine how much error is associated with a controversial psychological phenomenon we might want to measure such as intelligence. With classical psychometric true score theory, we can stop "imagining" how much error there is in our measurements and start estimating it.

Classical **true score theory** states that our observed score (X) is equal to the sum of our true score, or true underlying ability (T), plus the measurement error (E) associated with estimating our observed scores, or

$$X = T + E$$

Several assumptions are made about the relationship among these three components. These assumptions are discussed in detail in texts such as Allen

and Yen (1979) and Crocker and Algina (1986), so we will not cover them here. Briefly, however, the "true score" is the score we would obtain if we were to take the average score for an infinite number of test administrations. Of course, in practice, one cannot administer a test an infinite number of times, and as noted previously, the vast majority of the time we get only one chance. Therefore, we use **reliability coefficients** to estimate both true and error variance associated with our observed test scores. Theoretically speaking, our reliability estimate is the ratio of the true variance to the total variance:

$$r_{xx} = \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{total}}} = \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{true}} + \sigma^2_{\text{error}}}$$

where r_{xx} is the reliability, σ^2_{true} is the true score variance, σ^2_{total} is the total score variance, and σ^2_{error} is the error variance. Of course, we will never be able to directly estimate the true score and its variance; hence, this particular formula serves merely as a heuristic for understanding the components of reliability. In the following discussion, we will outline several options for estimating test reliability in practice.

Estimating Reliability in Practice

Right about now you are probably saying to yourself, "Okay, that's the theoretical stuff that my textbook talked about, but how do I actually compute a reliability estimate when I need to?" Most of the time, we compute a Pearson product moment correlation coefficient (correlation coefficient for short) or some other appropriate estimate (e.g., a Spearman correlation if we have ordinal data) to estimate the reliability of our measurement scale.

Before we can calculate our reliability estimate, however, we have to decide what type of measurement error we want to focus on. For example, if we want to account for changes in test scores due to time, we calculate the correlation coefficient between a test given at time 1 and the same test given at some later point (i.e., test-retest reliability). Therefore, once we know what source of error we want to focus on, we will know what type of reliability coefficient we are dealing with and thus which correlation coefficient to compute to estimate our reliability.

Looking at the first column of Table 5.1, you will notice three different sources of measurement error that we can estimate with different types of reliability estimates. Notice, however, that one source of measurement error,

content sampling, appears twice. Each of these sources can be considered to be tapping into the issue of how consistent our measures are. The first source of error, change in examinees, estimates how consistently our examinees respond from one occasion to another. The next two, content sampling, estimate how consistent items are across test versions or within a given test. Finally, we may estimate the consistency of raters' judgments of examinees or test items.

Table 5.1 Sources of Error and Their Associated Reliability and Statistics

Source of Error	*Reliability Coefficient*	*Reliability Estimate*	*Statistic*
Change in examinees	Stability	Test/retest	r_{12}
Content sampling	Equivalence	Alternate forms	$r_{xx'}$
Content sampling	Internal consistency	Split-half Alpha	r_{x1x2} α
Inter-rater	Rater consistency	Inter-rater	kappa

The second column in Table 5.1 lists the type of reliability coefficient to use when we estimate a given source of measurement error. For example, with changes in examinees our reliability coefficient is an estimate of stability or consistency of examinees' responses over time. With the first form of content sampling measurement error (i.e., row 2 in Table 5.1), we are measuring the equivalence or consistency of different forms of a test. However, with the second form of content sampling measurement error (i.e., row 3 in Table 5.1), we are measuring what is commonly referred to as the **internal consistency** of test items. That is, do all the items in a single test seem to be tapping the same underlying construct? Finally, when we estimate inter-rater sources of measurement error, our reliability coefficient is one of rater consistency. That is, do raters seem to be rating the target in a consistent manner, whether the target is the test itself or individuals taking the test, such as job applicants during an employment interview?

The third column in Table 5.1 presents the name of the reliability estimate we would use to determine the respective sources of measurement error. For example, if we wanted to estimate sources of measurement error associated

with changes in examinees over time, we would compute a **test-retest reliability** estimate. Alternatively, if we wanted to examine the first form of content sampling, we would compute what is commonly known as **alternate forms reliability** (also referred to as **parallel forms reliability**). However, for the second form of content sampling, we do not need to have a second form of the test. We would instead compute either a **split-half reliability coefficient** or an alpha reliability coefficient to estimate the internal consistency of a single test. Finally, for estimating sources of error associated with raters, we would compute an inter-rater reliability estimate.

You may remember that in Module 2 we discussed the importance of the correlation coefficient. Why is it so important? As you can see in the last column of Table 5.1, the correlation coefficient is used to compute most forms of reliability estimates. The only difference among the different correlation coefficients is the variables used and the interpretation of the resulting correlation. For example, with a test-retest reliability estimate, we would compute the correlation coefficient between individuals' scores on a given test taken at time 1 and those same individuals' scores on the same test taken at a later date. The higher the correlation coefficient, the more reliable the test or, conversely, the less error attributable to changes in the examinees over time. Of course, many things can affect the test-retest estimate of reliability. One is the nature of the construct being assessed. For example, when we are measuring enduring psychological traits, such as most forms of personality, there should be little change over time. However, if we were measuring transitory psychological states, such as fear, then we would expect to see more change from the first to the second testing. As a result, our test-retest reliability estimates tend to be lower when we are measuring transitory psychological states rather than enduring psychological traits.

In addition, the length of time between testing administrations can affect our estimate of stability. For instance, if we are measuring a cognitive skill such as fluency in a foreign language, and there is a long time period between testing sessions, an individual may be able to practice that language and acquire additional fluency in that language between testing sessions. As a result, individuals will score consistently higher on the second occasion. However, differences in scores from time 1 to time 2 will be interpreted as instability in subjects (i.e., a lot of measurement error) and not learning. On the other hand, if we make the duration between testing sessions too short, examinees may remember their previous responses. There may also be fatigue effects associated with the test-retest if the retest is immediate. So how long should the interval between testing sessions be? Unfortunately, there is no hard-and-fast rule. The key is to make sure that the duration is long enough not to fatigue the examinees or allow them to remember their answers, but

not so long that changes may take place (e.g., learning, psychological traumas) that could impact our estimate of reliability. Of course, one way to deal with the possibility of subjects remembering their answers from time 1 to time 2 is to use two different forms of the test.

With alternate forms reliability, we administer examinees one form of the test and then at a later date give them a second form of the test. Because we do not have to worry about the individuals remembering their answers, the intervening time between testing sessions does not need to be as long as with test-retest reliability estimates. In fact, the two testing sessions may even occur on the same day. From a practical standpoint, this may be ideal, in that examinees may be unwilling or simply fail to return for a second testing session. As you have probably surmised, the biggest disadvantage of the alternate forms method is that you need to have two versions of the test. It is hard enough to develop one psychometrically sound form of a test, now you have to create two. Is it possible to just look at content sampling within a single test?

With split-half and alpha reliability estimates, we need only one version of the test. To estimate split-half reliability, we correlate one half of the test with the other half of the test. If we simply correlate the first half with the second half, however, we may have spuriously low reliability estimates due to fatigue effects. In addition, many cognitive ability tests are spiral in nature, meaning they start out easy and get harder as you go along. As a result, correlating the first half of the test with the second half of the test may be misleading. Therefore, to estimate split-half reliability, most researchers correlate scores on the odd-numbered items with scores on the even-numbered items. As you might have guessed by now, we are, in a sense, computing a correlation on only half of our test. Does that in and of itself result in a lower reliability estimate? In fact, it does. Therefore, whenever a split-half reliability estimate is calculated, one should also use the **Spearman-Brown prophecy formula** to correct for the fact that we are cutting the test in half. (*Note:* We demonstrate an alternate use of the formula in Case Study 5.2.)

$$r_{XX'_n} = \frac{nr_{XX'}}{1 + (n - 1)r_{XX'}}$$

where $r_{XX'_n}$ is the Spearman-Brown corrected split-half reliability estimate; n is the factor by which we want to increase the test, which in this case would be 2 (because we are, in a sense, doubling the test back to its original length); and $r_{XX'}$ is the original split-half reliability estimate. Because n is always equal to 2 when correcting our split-half reliability estimate, our formula can be simplified to

$$r_{XX'_n} = \frac{2r_{XX'}}{1 + r_{XX'}}$$

The general form of the Spearman-Brown formula can be used to determine the estimated reliability of a revised version of the test if the number of items on the test is increased (or even decreased) by a specified factor. It is important to note, however, that the formula assumes that the additional items contributed to the test are parallel to the items on the original test. Thus, the new items must be similar to the original items in terms of content, difficulty, correlation with other items, and item variance.

The second, and more common, measure of internal consistency reliability is the alpha reliability estimate. **Coefficient alpha** is sometimes referred to as the average of all possible split-half reliabilities. As a result, the formula for computing alpha is a little more involved than a simple bivariate correlation coefficient:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right)$$

where α is the estimate of the alpha coefficient, k is the number of items on the test, σ_i^2 is the variance of item i, and σ_x^2 is the total variance of the test.

All other things being equal, the more items you have on your test (k) the higher your alpha coefficient will be. Hence, one way to increase the reliability of your test is to increase the number of items on the test. In addition, the alpha coefficient will also increase if we increase the variability of each item. Hence, removing items with very little variability from a test and replacing them with higher-variability items will actually increase your alpha coefficient.

How does one interpret the alpha coefficient? Actually, the interpretation is very similar to that of the other reliability estimates based on correlation coefficients. Zero would indicate no reliability (i.e., all measurement error). A value of one, on the other hand, would indicate perfect reliability (i.e., no measurement error). Thus, the common standard of a reliability estimate of at least .70 or higher holds for alpha as well.

Two precautions should be kept in mind when interpreting alpha reliability estimates. First, many students and practitioners often refer to the alpha coefficient as "the" estimate of reliability. As should be clear by now, based on our discussion of Table 5.1, the alpha coefficient is but one estimate of reliability that focuses on just one form of measurement error.

Therefore, if you are interested in other forms of measurement error (such as stability over time), you will need to compute additional reliability estimates. Second, as Cortina (1993) and Schmitt (1996) pointed out, one common misconception of alpha among naive researchers is that the alpha coefficient is an indication of the unidimensionality of a test. As pointed out previously, if you have a large enough set of items, you will have a high alpha coefficient, but this does not mean your test is unidimensional. The measurement of job satisfaction can serve as a good example of this phenomenon. Most job satisfaction scales measure several different facets of job satisfaction, such as satisfaction with one's job, supervisor, pay, advancement opportunities, and so on. However, the scales can also be combined to create an overall job satisfaction score. Clearly, this overall job satisfaction score is not unidimensional. Because the overall score is typically based on a large number of items, however, the overall scale's alpha coefficient will be large. As a result, it is important for researchers to remember that an alpha coefficient only measures one form of measurement error and is an indication of internal consistency, not unidimensionality.

Finally, to estimate inter-rater agreement, a statistic such as Cohen's kappa can be used. To compute kappa, sometimes referred to as scorer reliability, you would need to set up a cross-tabulation of ratings given by raters, similar to a chi-square contingency table. For example, you might have a group of parents, both the mother and the father (your two raters), rate their children on the children's temperament (e.g., 1 = easygoing, 2 = anxious, 3 = neither). You would want to then determine if the parents agree in terms of their respective perceptions (and ratings) of their children's temperaments. To compute the **kappa statistic,** you would need to set up a 2(raters) × 3(temperament rating) contingency table of the parents' ratings. Then you would compute the kappa statistic as follows:

$$k = \frac{(Oa - Ea)}{(N - Ea)}$$

where k is the kappa statistic, Oa is the observed count of agreement (typically reported in the diagonal of the table), Ea is the expected count of agreement, and N is the total number of respondent pairs. Thus, Cohen's kappa represents the proportion of agreement among raters after chance agreement has been factored out. In this case, zero represents chance ratings, while a score of one represents perfect agreement. (*Note:* Exercise 5.3 provides data for computing Cohen's kappa.)

As with many statistics, however, kappa has not been without its critics (e.g., Maclure & Willett, 1987). One criticism is that kappa is not a good estimate of effect size. Although it will give a pretty good estimate of whether the observed ratings are significantly different from chance (an inferential statistic), using kappa as an estimate of the actual degree of agreement (i.e., as an effect size estimate) should be done cautiously, as the statistic assumes the raters are independent. In our preceding example, it is highly unlikely that the parents will provide independent ratings. Thus, when it can be reasonably assumed that raters are not independent, you would be better off using other estimates of rater agreement, such as the intraclass correlation coefficient.

Thus, we see there are many forms of reliability, each of which estimates a different source of measurement error. In general, immediate test-retest reliability and split-half reliability tend to provide upper-bound estimates of reliability. That is, they tend to provide higher estimates, on average, than other forms of reliability. Coefficient alpha and long-term test-retest tend to provide somewhat lower estimates, on average, while alternate forms reliability, both short and long term, tends to provide lower-bound estimates. Why present this information here? These general trends are important both for interpreting your obtained reliability coefficients and for using your reliability estimates for other purposes, such as determining the standard error of measurement.

What Do We Do With the Reliability Estimates Now That We Have Them?

You are probably asking yourself, "Now that we have an estimate of reliability, what do we do with it?" First, we will need to report our reliability estimate(s) in any manuscripts (e.g., **technical manuals,** conference papers, and articles) that we write. Second, if we have followed sound basic test construction principles, someone who scores high on our test is likely to be higher on the underlying trait than someone who scores low on our test. Often times, this general ranking is all we are really looking for; who is "highest" on a given measure. However, if we want to know how much error is associated with a given test score (such as when we set standards or cutoff scores), we can use our reliability estimate to calculate the **standard error of measurement,** or SEM (of course, we would also need to know the sample standard deviation for the measure). Thus, computing the SEM allows us to build a confidence interval around our observed score so that we can estimate (with a certain level of confidence) someone's underlying true score,

$$SEM = S_x \sqrt{1 - r_{xx}}$$

where S_x is the sample standard deviation and r_{xx} is the reliability estimate.

EXAMPLE: $X = 100$, $S_x = 10$, $r_{xx} = .71$

$$SEM = 10 \sqrt{1 - .71} = 10\ (.5385) = 5.38$$

$$95\%\ \text{CI} = X \pm 1.96^{*}\text{SEM} = 100 \pm 1.96^{*}(5.38)$$
$$= 100 \pm 10.54 = 89.46 \leq T \leq 110.54$$

where X is our test score, 1.96 is the critical z value associated with the 95% confidence interval, SEM is the standard error of measurement value, and T is our estimated underlying true score value.

You can see from the preceding formula that, as our test becomes more reliable, our confidence interval becomes narrower. For example, if we increase the reliability of our test to .80, the SEM in the previous example becomes 4.47 and thus the 95% confidence interval narrows to $91.24 \leq T \leq 108.76$. We could even reverse the formula and figure out how reliable our test needs to be if we want a certain width confidence interval for a test with a given standard deviation. For example, if we want to be 95% confident that a given true score is within 5 points (SEM = 2.5, plus or minus in either direction) of someone's observed score, then we would have to have a test with a reliability of .9375:

$$SEM = S_x \sqrt{1 - r_{xx}},\ \textit{becomes}\ 1 - \left[\frac{SEM}{S_x}\right]^2 = 1 - \left[\frac{2.5}{10}\right]^2 = 1 - .0625 = .9375$$

Concluding Comments

There will always be some degree of error when we try to measure something. Physical characteristics, however, tend to have less measurement error than psychological phenomena. Therefore, it is critical that we accurately estimate the amount of error associated with any measure, in particular, psychological measures. To estimate the measurement error, we have to first decide what form of error we are most interested in estimating. Once we do

that, we can choose an appropriate reliability estimate (see Table 5.1) to estimate the reliability. We can then use the reliability estimate to build confidence intervals around our observed scores to estimate the underlying true scores. In doing so, we will have much more confidence in the interpretation of our measurement instruments.

Practical Questions

1. How much can we shorten an existing measure and still maintain adequate reliability? (See Case Study 5.2.)
2. What are the different sources of error that can be assessed with classical reliability analysis?
3. Does it matter which reliability estimate we put into the standard error of measurement formula?
4. Are some reliability estimates generally higher (or lower) than others? That is, does one tend to serve as an upper- (or lower-) bound reliability estimate?
5. How is Cohen's kappa estimate of reliability different from the other forms of reliability?
6. Why are some authors (e.g., Cortina, 1993; Schmitt, 1996) cautious about the interpretation of coefficient alpha?

Case Studies

CASE STUDY 5.1: DON'T FORGET TO REVERSE SCORE

It didn't make sense. It just didn't. How could the reliability be so low? Chad scratched his head and thought. Chad had agreed to help analyze the data from his graduate advisor's most recent study. Although entering the data into a computer database had not been exciting, it had been relatively easy. Once he had entered each research participant's responses, he spot-checked a few cases to ensure accuracy. He then conducted frequency analyses on each variable to ensure that there were no out-of-bounds responders. In fact, he'd found two cases in which he had incorrectly entered the data. He

could tell, because items that were responded to on a five-point Likert-type rating scale had reported scores of 12 and 35, respectively. Sure enough, he'd just made a typo when entering the data. Everything else looked fine.

Or so he thought, until he decided to examine the reliability of one of the scales. Chad's advisor, Dr. John Colman, was primarily interested in troubled adolescents, and over the last several years had investigated adolescent attitudes toward alcoholic beverages. The same measure of adolescent attitudes toward alcohol was routinely used in this research. Respondents indicated on a scale of 1–5 how strongly they agreed with each of the 12 items. Internal consistency reliability estimates for the scale were consistently good, typically around .80. However, not this time, apparently. In computing the reliability estimate for the data he'd just entered, Chad found that alpha was estimated to be –.39.

Chad couldn't remember ever hearing of a negative internal consistency reliability estimate. In addition, he couldn't explain why the scale would have such a different reliability on this sample than it had with the many samples his advisor had previously used. His first thought was that he might have entered the data incorrectly—but he knew he hadn't. After all, he'd checked the data carefully to ensure that the computer data file matched exactly what was on the original surveys. So what could be the problem?

In examining the item-total correlations for each item on the scale, Chad noticed that several items correlated negatively with a composite of the remaining items. Chad grabbed the original survey and reexamined the 12 items that comprised the adolescent attitudes toward alcohol scale. Each item certainly seemed to measure the intended construct. Chad was about to give up and go report the problem to his advisor when he noticed something. Although each of the 12 items measured attitudes toward alcohol, agreement to eight of the items would be indicative of acceptance of alcohol use. In contrast, agreement to the other four items would be indicative of a rejection of alcohol use. That was it. He'd correctly entered the data from the surveys into the computer data file, but had forgotten to recode the reverse-coded items. Because his advisor wanted high scores to be indicative of an acceptance of the use of alcohol, Chad decided he'd recode the four reverse-coded items. To do this, he used the recode command of his statistics program to recode all responses of "5" into "1," "4" into

"2," "2" into "4," and "1" into "5." He did this for each of the four reverse-coded items. Holding his breath, he again computed the alpha. This time, the reliability estimate was $\alpha = .79$, and all of the item-total correlations were positive. Satisfied that he'd been able to resolve the problem on his own, Chad made a mental note to always recode the appropriate items once the entire data file had been completed.

Questions to Ponder

1. In terms of Table 5.1, what type of reliability coefficient did Chad estimate? What source of error is being estimated?
2. Did Chad make the right interpretation of his negative reliability estimate? What else might cause a negative reliability estimate?
3. In practice, how does one know which items to recode and which to keep the same?
4. Both positively and negatively worded items are frequently included on tests. Assuming you recode the negatively worded items before you run your reliability analysis, will the inclusion of negatively worded items affect the test's internal consistency reliability estimate?

CASE STUDY 5.2: LENGTHENING AND SHORTENING PSYCHOLOGICAL SCALES

Sheila was frustrated. Although she was happy with both the topic and the constructs she had chosen to examine in her senior honors thesis, she had hit several roadblocks in determining what measures to use to assess each variable in her proposed study. Now that she had finally identified useful measures to include in her survey, she was concerned that her response rate would suffer because of the rather impressive length of the survey. Reasoning that the sample she hoped to use was unlikely to spend more than a few minutes voluntarily responding to a survey, Sheila considered her options. First, she could eliminate one or more variables. This would make her study simpler and would have the added benefit of reducing the length of the survey. Sheila rejected this option, however, because she felt each variable she had identified was necessary to adequately address her research questions. Second, she considered just mailing the survey to a larger number of people in order to get an adequate number to respond to the lengthy survey. Sheila quickly rejected this option as well. She certainly didn't want to pay for the additional copying and mailing costs. She was also

concerned that a lengthy survey would further reduce the possibility of obtaining a sample that was representative of the population. Perhaps those individuals who would not respond to a long survey would be very different from the actual respondents.

Suddenly a grin spread across Sheila's face. "Couldn't I shorten the survey by reducing the number of items used to assess some of the variables?" she thought. Some of the scales she had selected to measure variables were relatively short, while scales to measure other variables were quite long. Some of the scales were publisher-owned measures and thus copyrighted. Others were nonproprietary scales both created and used by researchers. Recognizing the reluctance of publishers to allow unnecessary changes to their scales, Sheila considered the nonproprietary measures. The scale intended to assess optimism was not only nonproprietary but also very long: 66 items. A scale assessing dogmatism was also nonproprietary and, at 50 items, also seemed long. Sheila quickly decided that these would be good scales to target for reduction of the number of items.

In class, Sheila had learned that the Spearman-Brown prophecy formula could be used to estimate the reliability of a scale if the scale was doubled in length. Her instructor also explained that the same formula could be used for either increasing or decreasing the number of items by a certain factor. Sheila knew from her research that the typical internal consistency reliability finding for her optimism scale was .85, and for the dogmatism scale it was .90. Because she wanted to reduce the number of items administered for each scale, she knew the resulting reliability estimates would be lower. But how much lower? Sheila considered reducing the number of items in both scales by one half. Because she was reducing the number of items, the number of times she was increasing the scale was equal to one half, or .5. She used this information to compute the Spearman-Brown reliability estimate as follows:

Optimism Test	Dogmatism Test
$r_{XX'_n} = \frac{nr_{XX'}}{1 + (n-1)r_{XX'}}$	$r_{XX'_n} = \frac{nr_{XX'}}{1 + (n-1)r_{XX'}}$
$= \frac{.5(.85)}{1 + (.5-1).85}$	$= \frac{.5(.90)}{1 + (.5-1).90}$
$= .74$	$= .82$

In considering these results, Sheila thought she'd be satisfied with an internal consistency reliability estimate of .82 for the dogmatism scale, but was concerned that too much error would be included in estimates of optimism if the internal consistency reliability estimate were merely .74.

Undeterred, Sheila decided to estimate the reliability if only one third of the optimism items were removed. If one third of the items were dropped, two thirds (or .67) of the original items would remain. Therefore, the Spearman-Brown prophecy estimate could be computed as follows:

Optimism Test

$$r_{XX'_n} = \frac{nr_{XX'}}{1 + (n - 1)r_{XX'}}$$

$$= \frac{.67(.85)}{1 + (.67 - 1).85}$$

$$= .79$$

Sheila decided this reliability would be acceptable for her study. In order to complete her work, Sheila randomly selected 25 (50%) of the items from the dogmatism scale, and 44 (67%) of the items from the optimism scale. She was confident that although her survey form was now shorter, the reliability of the individual variables would be acceptable.

Questions to Ponder

1. In terms of Table 5.1, what type of reliability coefficient did Sheila estimate? What source of error is being estimated?
2. Should Sheila have randomly selected which items to keep and which to delete? What other options did she have?
3. How else might Sheila maintain her reliability levels yet still maintain (or increase) the number of usable responses she obtains?
4. Why do you think Sheila is using .80 as her lower acceptable bound for reliability?

Exercises

EXERCISE 5.1: COMPUTING TEST-RETEST, ALPHA, AND PARALLEL FORMS RELIABILITY VIA COMPUTER

OBJECTIVE: To practice calculating different types of reliability.

Using the data set "Reliability.sav" (see the variable list in Appendix B), perform the reliability analyses outlined below. The scales provided here include a depression scale (14 items, V1–V14), a life satisfaction scale (10 items, V15–V24), a reasons-a-person-retired scale (10 items, V25–V34), a scale with regard to good things about retirement (8 items, V35–V42), and a scale with regard to bad things about retirement (6 items, V43–V48). For your assignment (be sure to do an ocular analysis of all items first, checking for outliers, missing data, etc., before jumping into the reliability analyses):

1. Perform alpha, split-half, and parallel forms reliability analyses for each of the five scales. How do the three different types of reliability compare for each scale listed above? Is one form of reliability more appropriate than another? Discuss for each scale. (*Note*: You may wish to put your results in table form for easy comparison.)

2. Using alpha reliability, with item and scale information, what items should be included in the final versions of each scale in order to maximize the alpha reliability for that scale? (*Note:* You will need to examine the item-total correlations. In addition, once an item is removed, you will need to repeat the process until a final scale is decided upon.)

3. For the life satisfaction and depression scales, determine if the alpha reliabilities are different for men and women (SEX). If yes, any guesses why? (*Note:* This requires using the "split file" option in SPSS or comparable options in other statistics programs.)

4. Based on Cortina (1993), what additional analyses could you conduct to evaluate the reliabilities of each of the scales? (You may perform these analyses if you wish, but it is not required for this assignment.)

EXERCISE 5.2: EXAMINING THE EFFECTS OF THE SPEARMAN-BROWN PROPHECY FORMULA

OBJECTIVE: To practice using the Spearman-Brown prophecy formula for estimating reliability levels.

Using the Spearman-Brown prophecy formula provided in Case Study 5.2, estimate Sheila's reliability for the dogmatism scale if she used only one third of the number of original items. Is this an "acceptable level" of reliability? Why or why not?

EXERCISE 5.3: ESTIMATING AGREEMENT COEFFICIENTS (COHEN'S KAPPA)

OBJECTIVE: To practice calculating Cohen's kappa estimate of rater agreement.

Assume you wanted to determine the degree of inter-rater agreement between two forensic psychologists who were each rating 100 potential parolees in terms of their potential for committing additional violent crimes. In general, sociopaths are more likely to commit additional violent crimes than are depressed or normal individuals. Therefore, each psychologist rated each of the 100 potential parolees on a scale of 1–3 in terms of their primary personality category (1 = sociopath, 2 = depressed, 3 = normal). The following results were obtained:

		Forensic Psychologist A		
Forensic		Personality 1	Personality 2	Personality 3
Psychologist B	Personality 1	44	5	1
	Personality 2	7	20	3
	Personality 3	9	5	6

Using the data in the preceding table and the formula for kappa presented in the module overview, determine the level of agreement between the raters.

Internet Web Site References

5.1. http://trochim.human.cornell.edu/kb/reliable.htm

This Web page presents the beginning of the chapter on reliability from Dr. William M. Trochim's electronic textbook.

5.2. http://trochim.human.cornell.edu/tutorial/levell/mazeintr.htm

This Web page provides a 10-item quiz on reliability, along with a rationale for the correct answer to each item.

5.3. http://www.statsoftinc.com/textbook/streliab.html

This Web page presents the beginning of the chapter on reliability from Statsoft, Inc.'s electronic textbook.

5.4. http://www.ruf.rice.edu/~lane/stat_sim/reliability_reg/

This Web page presents a simulation of the effects of the reliability of *X* and *Y* on a number of components of regression analysis.

5.5. http://www.unl.edu/BIACO/workshops/reliability%20folder/sld001.htm

This Web page is the start page of a slide presentation by the Buros Institute for Assessment Consultation and Outreach (BIACO) titled "Consistency in Scoring (Reliability) Workshop."

5.6. http://chiron.valdosta.edu/mawhatley/3900/reliablec.htm

This Web page provides a 30-item reliability quiz. Responses are scored immediately.
Note: Because of the overlap between the concepts presented in Modules 6–8, many of the links listed under a particular module in this section are likely to present additional information relevant to the other modules.

Further Readings

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Maclure, M., & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology, 126,* 161–169.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and application.* Thousand Oaks, CA: Sage.