

CHAPTER FIVE

Investigating reliability for norm-referenced tests

II. Conceptual exercises

A. Matching

1. g
2. e
3. a
4. j
5. b
6. d
7. i
8. c
9. f
10. h

B. True or false

1. T
2. F
3. T
4. F
5. T
6. T

C. Brief answers

1. Some potential sources of variation in scores from an oral interview:

Language ability: knowledge of grammar, vocabulary, cohesion, conversational organization, register, metacognitive strategies

Personal characteristics: personality (outgoing or shy), willingness to take risks, L2 ability

Test method: prompts that are presented, the content domain of the prompts, the characteristics of the interviewers (personality, age, gender and status), the quality of the tape recording, the way the rater interprets and applies the rating scale

Random: how the test taker feels on the day of the interview, the way examinees are paired with interviewers, the specific prompts the interviewers choose to use

2. Explain why we need to assume that the two sets of scores used in classical reliability estimates are independent of each other.

If one score is affected by the other, then we cannot tell whether performance is due to the ability we want to measure or whether it's because of the dependency between the scores. That is, we won't be able to unambiguously interpret the correlation between the two sets of scores as being due to the test takers' true scores and not to the effect of one score on the other. Dependence between the two sets of scores will lead to overestimation of correlation between them and thus overestimation of the reliability estimates.

3. What three general steps are followed in the procedures for calculating all classical reliability coefficients?
 - a. identifying the relevant source of measurement error;
 - b. designing a study to collect two sets of independent and parallel scores; and
 - c. estimating the reliability by calculating either the appropriate correlation coefficient or coefficient alpha.
4. Suppose you have given a reading test to a group of test takers. This test consists of two reading passages, each followed by ten multiple-choice items. What is the most appropriate estimate of the internal consistency of scores from this test, and why?

Since the questions that follow each passage may all depend on the same reading passage, we cannot assume that they are independent. Furthermore, depending on how the items are ordered, answering some questions correctly may help test takers answer others correctly. Because of this dependency, the most appropriate approach would be to use a G-study design with items nested within passages. The design could be represented as $p \times i:pa$, where 'p' is persons, 'I' is items and 'pa' is passage. In this design, there would be two conditions for the passage facet and ten conditions for the items facet.

5. Explain why it is important to use a counterbalanced design when collecting data to estimate equivalent forms reliability.

Because it is not possible for test takers to take two forms of the test at the same time, the tests must be administered in order, with one first and the other second. We might expect that test takers will do better on the second form, because of the practice effect. To cancel out this practice affect, we give half the test takers one test first and the other half the other form first.

6. Explain how the standard error of measurement (SEM) can be used to estimate the reliability of a given test taker's score.

When we calculate the SEM, we are estimating how much error variance there is, on average, in the distribution of test scores. Since the SEM is scaled on the same scale as the raw scores, we can use this to calculate a confidence interval within which we believe the test taker's true score falls. We can use different levels of confidence, depending on the importance of the decisions to be made on the basis of the test. If it's a high-stakes test, then we would probably use the .99 CI. If it's a low-stakes test, we could use a lower CI, say, .95

7. Suppose you wanted to use G-theory to estimate the dependability of the scores in the oral interview assessment described in exercise II. C. 1 above. What would the facets of your design be?

The facets of the measurement procedure would be prompts, interviewers and scorers.

8. If this assessment were to be used to select applicants for admission to an academic program, what estimate of dependability would be appropriate?

This will depend on whether it's a relative or absolute decision. Assuming that this is a relative decision, since the number of places in the program is likely to be limited, and we want to admit the top students, the generalizability coefficient, ρ , would be the most appropriate. If we wanted to estimate the consistency of decisions at a particular cut score, then the ϕ lambda would be appropriate.

9. If this assessment were to be used to certify test takers as competent in academic oral language, what estimate of dependability would be appropriate?

Since this would be an absolute decision, and we'd want to estimate the decision consistency at a particular cut score, phi lambda would be appropriate.

III. Hand calculations with small data sets

A. Using the scores below, calculate the split-half reliability, using the Spearman-Brown correction formula (answer dataset: rel-split-half1-ans.doc on the CD). Then calculate the Guttman split-half for the same data set.

Student ID	X_T	X_O	X_e	R_O	R_e	d $R_e - R_O$	d^2
1.	29	15	14	8.5	13	4.5	20.25
2.	33	16	17	5	9	4	16
3.	32	14	18	12	5.5	6.5	42.25
4.	36	15	21	8.5	2	6.5	42.25
5.	34	16	18	5	5.5	0.5	0.25
6.	32	15	17	8.5	9	0.5	0.25
7.	17	7	10	15	14.5	0.5	0.25
8	35	18	17	2.5	9	6.5	42.25
9.	31	14	17	12	9	3	9
10.	36	16	20	5	3.5	1.5	2.25
11.	37	20	17	1	9	8	64
12.	40	18	22	2.5	1	1.5	2.25
13.	32	12	20	14	3.5	10.5	110.25
14.	31	15	16	8.5	12	3.5	12.25
15.	24	14	10	12	14.5	2.5	6.25
$\bar{X} =$	31.9	15.0	16.9				
$s^2 =$	31.07	8.71	12.07				$\Sigma d^2 = 370.00$

Here are the steps to follow.

Spearman correlation between halves:

$$r_{hh'}(\rho) = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

$$r_{hh'} = 1 - \frac{6(370)}{15(225 - 1)}$$

$$r_{hh'} = 1 - .661$$

$$r_{hh'} = .339$$

Spearman-Brown correction:

$$r_{xx'} = \frac{2r_{hh'}}{1 + r_{hh'}}$$

$$r_{xx'} = \frac{2(.339)}{1 + .339}$$

$$r_{xx'} = \frac{.678}{1.339}$$

$$r_{xx'} = .506$$

Guttman split-half:

$$r_{xx'} = 2 \left(1 - \frac{S^2_{h_1} + S^2_{h_2}}{S^2_x} \right)$$

$$r_{xx'} = 2 \left(1 - \frac{8.71 + 12.07}{31.07} \right)$$

$$r_{xx'} = 2 (1 - .6688)$$

$$r_{xx'} = .662$$

B. Using the scores below, calculate the split-half reliability, using the Spearman-Brown correction formula (answer data set: rel-split-half2-ans.doc on the CD). Then calculate the Guttman split-half for the same data set.

Student ID	X _T	X _O	X _e	R _O	R _e	d R _e -R _O	d ²
1.	22	10	12	14	14.5	.5	.25
2.	31	15	16	7.5	8	.5	.25
3.	32	15	17	7.5	6	1.5	2.25
4.	38	20	18	3	3.5	.5	.25
5.	32	17	15	5	10	5	25
6.	31	15	16	7.5	8	.5	.25
7.	23	10	13	14	12.5	1.5	2.25
8	37	19	18	4	3.5	.5	.25
9.	26	10	16	14	8	6	36
10.	32	14	18	10	3.5	6.5	42.25
11.	40	22	18	1	3.5	2.5	6.25
12.	45	21	24	2	1	1	1
13.	29	15	14	7.5	11	3.5	12.25
14.	25	13	12	11	14.5	3.5	12.25
15.	24	11	13	12	12.5	.5	.25
$\bar{X} =$	31.1	15.1	16.0				
$s^2 =$	44.52	16.12	9.71				$\Sigma d^2 = 141.00$

Here are the steps to follow.

Spearman correlation between halves:

$$r_{hh'}(\rho) = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$

$$r_{hh'} = 1 - \frac{6(141)}{15(225 - 1)}$$

Spearman-Brown correction:

$$r_{xx'} = \frac{2r_{hh'}}{1 + r_{hh'}}$$

$$r_{xx'} = \frac{2(.748)}{1 + .748}$$

$$r_{hh'} = 1 - .252 \qquad r_{xx'} = \frac{1.496}{1.748}$$

$$r_{hh'} = .748 \qquad r_{xx'} = .856$$

Guttman split-half:

$$r_{xx'} = 2 \left(1 - \frac{S_{h_1}^2 + S_{h_2}^2}{S_x^2} \right)$$

$$r_{xx'} = 2 \left(1 - \frac{16.12 + 9.71}{44.52} \right)$$

$$r_{xx'} = 2 (1 - .5802)$$

$$r_{xx'} = .840$$

C. Using the data below, calculate the reliability of the total scores, using the formula for coefficient alpha (answer data set: relrater-ans.doc on the CD).

Descriptive Statistics

	N	Mean	Variance
	Statistic	Statistic	Statistic
RATER1	54	3.2963	.439
RATER2	54	3.2963	.439
RATER3	54	3.2593	.460
TOTAL	54	9.8519	3.110
Valid N (listwise)	54		

$$\alpha = \frac{k}{k-1} \left(1 - \frac{s_{r1}^2 + s_{r2}^2 + s_{r3}^2}{s_T^2} \right) \qquad \alpha = 1.5(1 - .430)$$

$$\alpha = \frac{3}{3-1} \left(1 - \frac{.439 + .439 + .460}{3.110} \right) \qquad \alpha = 1.5(.570)$$

$$\alpha = 1.5 \left(1 - \frac{1.338}{3.110} \right) \qquad \alpha = .855$$

V. SPSS exercises

A. Using the SPSS output for the Listening section of ESLPE F96 given above, fill in the values in the following table:

Coefficient	Value
Alpha (total)	.7234
1 st -2 nd half split, Spearman-Brown	.6993
1 st -2 nd half split, Guttman	.6990
Odd-even half split, Spearman-Brown	.7174
Odd-even half split, Guttman	.7166

B. Compare the differences among the different estimates of internal consistency reliability for the ESLPE F96 Listening section, and discuss why these might be different.

The coefficient alpha is slightly different from the split-half estimates because coefficient alpha treats individual items as parallel measures and bases the estimate on item variances whereas the split-half estimates are based on the corrected correlation between the scores on the halves. If the two halves are not statistically parallel, this would also explain why the split-half estimates are lower than coefficient alpha. The differences between the different split-half estimates are really too small to interpret.

- C. Why might none of the split-half estimates provided by these analyses be appropriate?

If the Listening section of this test is speeded, then none of these estimates would be appropriate. That is, if most of the test takers had difficulty comprehending the lecture because it was too fast, then these estimates would not be appropriate. Or if all of the Listening items are based on the same listening passage, none of these split-half estimates would be appropriate since the two halves would not be independent of each other. All the estimates would be overestimates.

- D. Using the ESLPE F96 data set, use the SPSS RELIABILITY procedure to calculate the following reliability estimates. You may either modify the commands in the SPSS syntax file, ESLPE F96-rel.sps, or use the pull-down menus in SPSS to do this.

Using the SPSS output for the Reading section of ESLPE F96, fill in the values in the following table:

Coefficient	Value
Alpha (total)	.8457
1 st -2 nd half split, Spearman-Brown	.7648
1 st -2 nd half split, Guttman	.7543
Odd-even half split, Spearman-Brown	.8421
Odd-even half split, Guttman	.8421

- E. Compare the differences among the different estimates of internal consistency reliability for the ESLPE F96 Reading section, and discuss why these might be different.

As with the Listening section, coefficient alpha is slightly different from the split-half estimates because coefficient alpha treats individual items as parallel measures and bases the estimate on item variances whereas the split-half estimates are based on the corrected correlation between the scores on the halves. In addition, if the two halves are not statistically parallel, then the Spearman-Brown split-half estimates underestimate the reliability. The Guttman split-half estimate is robust to violations

of parallelism. The difference between the Spearman-Brown and Guttman estimates based on the 1st-2nd half split suggests that the two halves are not parallel.

- F. Are any of the split-half reliability estimates for the reading section appropriate? Why or why not?

Since all the items in this section of the test are based on a single passage, we might expect there to be inter-item dependency. In this case, none of these estimates would be appropriate. A more appropriate estimate would be to design a study with two passages and then conduct a G-study with items nested within passages.