Shiken: JALT Testing & Evaluation SIG Newsletter, 4 (3) January 2001 (p. 7 - 11)

Statistics Corner

Questions and answers about language testing statistics:

Can we use the Spearman-Brown prophecy formula to defend low reliability?

James Dean Brown (University of Hawai'i at Manoa)

QUESTION: Can we defend a low (or undesirable) reliability coefficient by estimating the ideal number of test items with the Spearman-Brown prophecy formula? Hirai (1999) used the Spearman-Brown prophecy formula to defend her test's moderate reliability this way: with two sets of 8 MC [multiple-choice] questions, her Cronbach alpha was .70, which is moderate. This is hardly surprising considering the small number of items. If the standard Spearman-Brown prophecy formula for 30 items is applied to this test, however, the Cronbach's alpha estimated jumps to .81. (p. 375). So, her point may be that the reliability of .70 is good for a small number of test items. In other words, for a 30-item test, .81 is enough: therefore, for 16 items, .70 is adequate.

I have never seen the formula used like this. Hirai indicated that she got a hint from your book [Brown, 1996; or Brown with Wada, 1999] regarding this. To my knowledge, no one else has adapted the Spearman-Brown formula as she did. How common is this practice?

I know that Spearman-Brown formula is used as a split-half method to double a reliability coefficient after splitting the test into two; however, I didn't know that this formula might also be applicable to more than double the reliability coefficient for estimation.

(This question was submitted by Innami Yo of the MA program in TEFL at the University of Tsukuba.)

ANSWER: I find three separate questions in what you wrote above, two that you yourself posed in your first and second paragraphs and a third that is implied in your third paragraph: (a) Can we defend a low (or undesired level of) reliability coefficient by estimating the ideal number of test items on the Spearman-Brown prophecy formula? (b) Is this technique commonly used? (c) Can this formula be applied to more than double the reliability coefficient for estimation?

Can we defend a low (or undesired level of) reliability coefficient by estimating the ideal number of test items on the Spearman-Brown prophecy formula?

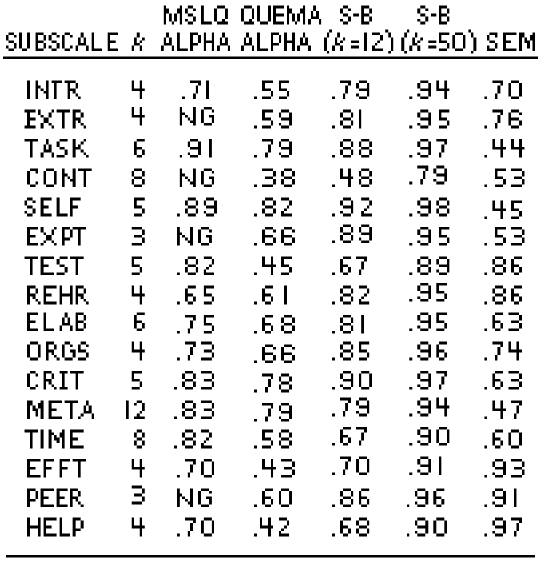
Hirai (1999) does indeed use the Spearman-Brown prophecy formula in discussing the reliability of her multiple-choice questions. The fact, that you and other readers may interpret this use of the Spearman-Brown formula as a defense of her "moderate" reliability points to one danger in making this kind of argument. No, we cannot defend a low reliability in that way. However, since

Hirai presents the original Cronbach reliability estimate of .70, I cannot fault her for making that argument. She was open and honest in her approach and can be said to be simply demonstrating that the reason for her "moderate" reliability was probably the relatively small number of items. The reader is still free to interpret the .70 she reported as indicating 70% reliability and 30% error.

Is this technique commonly used?

I wouldn't say that this strategy is commonly used, but it is sometimes used to make demonstrations of the sort Hirai did. I myself have used it several times. Most recently, I used it in Brown, Cunha, and Frota (2001), where we were examining the reliability of the subscales (of various lengths) on the QEMA (a Portuguese versions of the Motivated Strategies for Learning Questionnaire, or MSLQ):

Table 1. *Subscale and Cloze Reliabilities for a Portugese Learning Questionnaire.*
(Adapted from Brown, Cunha, and Frota, 2001)



		MSLQ	QUEMA	S-B	S-B	
SUBSCALE	*k*	ALPHA	ALPHA	(*k*=12)	(*k*=50)	SEM
INTR	4	.71	.55	.79	.94	.70
EXTR	4	NG	.59	.81	.95	.76
TASK	6	.91	.79	.88	.97	.44
CONT	8	NG	.38	.48	.79	.53
SELF	5	.89	.82	.92	.98	.45
EXPT	3	NG	.66	.89	.95	.53
TEST	5	.82	.45	.67	.89	.86
REHR	4	.65	.61	.82	.95	.86
ELAB	6	.75	.68	.81	.95	.63
ORGS	4	.73	.66	.85	.96	.74
CRIT	5	.83	.78	.90	.97	.63
META	12	.83	.79	.79	.94	.47
TIME	8	.82	.58	.67	.90	.60
EFFT	4	.70	.43	.70	.91	.93
PEER	3	NG	.60	.86	.96	.91
HELP	4	.70	.42	.68	.90	.97

To illustrate the relationship between the number of items and reliability, columns five and six of Table 1 give estimates of what the reliability would be if the subscales were all 12 items in length and 50 items in length, respectively. These adjustments were calculated using the Spearman-Brown prophecy formula (S-B). Notice that the reliabilities for each of the subscales are predictably higher for all of the subscales when adjusted to 12 items and are even very high when adjusted to 50 items. However, since such long subscales are clearly impractical, the demonstration here is simply meant to be illustrative and is not meant to suggest that the subscales actually be lengthened to 12 or 50 items.

Since Brown, Cunha, and Frota (2001) did include the original reliability estimates and since they, like Hirai, were simply demonstrating what the reliability would have been under various conditions, I personally think they were within logical and ethical bounds. However, I may not be an impartial judge in this case.

Can this formula be applied to more than double the reliability coefficient for estimation?

As you mentioned the Spearman-Brown prophecy formula is commonly used for adjusting split-half reliability estimates for full test reliability. To review briefly, split-half reliability is an internal consistency estimate. Split-half reliability is typically calculated in the following steps:

1. Divide whatever test you are analyzing into two halves and score them separately (usually the odd numbered items are scored separately from the even-numbered items).

2. Calculate a Pearson product-moment correlation coefficient between the students' scores on the even-numbered items and their scores on the odd-numbered items. The resulting coefficient is an estimate of the half-test reliability of your test (i.e., the reliability of the odd-numbered items, or the even-numbered items, but not both combined).

3. Apply the Spearman-Brown prophecy formula to adjust the half-test reliability to full-test reliability. We know that, all other factors being held constant, a longer test will probably be more reliable than a shorter test. The Spearman-Brown prophecy formula was developed to estimate the change in reliability for different numbers of items. The Spearman-Brown formula that is often applied in the split-half adjustment is as follows:

$$reliability = \frac{2 \times r_{half-test}}{1 + r_{half-test}}$$

For example, if the half-test correlation (for a 30-item test) between the 15 odd-numbered and 15 even-numbered items on a test turned out to be .50, the full-test (30-item) reliability would be .67 as follows:

$$reliability = \frac{2 \times r_{half-test}}{1 + r_{half-test}} = \frac{2 \times .50}{1 + .50} = \frac{1.00}{1.50} = .666 \approx .67$$

However, there is another version of the formula, which can be applied to situations other than a simple doubling of the number of items:

$$reliability = \frac{n \times r}{1 + (n-1)r}$$

Using the more complex formula, we get the same answer as we did with the simpler formula for the split-half reliability adjustment example as follows:

$$reliability = \frac{n \times r}{1+(n-1)r} = \frac{4.2 \times .50}{1+(4.2-1).50} = \frac{4.2 \times .50}{1+(3.2).50} = \frac{4.2 \times .50}{1+1.60} = \frac{2.10}{2.60} = .807 \approx .81$$

We can also use the more complex formula to estimate what the reliability for that same test would be if it had 60 items by using n = 4 (for the number of times we must multiply 15 to get 60; 4 x 15 = 60) as follows:

$$reliability = \frac{n \times r}{1+(n-1)r} = \frac{.33 \times .50}{1+(.33-1).50} = \frac{.33 \times .50}{1+(-.67).50} = \frac{.33 \times .50}{1+(-.335)} = \frac{.165}{.665} = .248 \approx .25$$

Or we can estimate what the reliability would be for various fractions of the test length. For instance, we could estimate the reliability for a 63 item test by using n = 4.2 (for the number of times we must multiply 15 to get 63; 4.2 x 15 = 63) as follows:

$$reliability = \frac{n \times r}{1+(n-1)r} = \frac{4 \times .50}{1+(4-1).50} = \frac{4 \times .50}{1+(3).50} = \frac{4 \times .50}{1+1.50} = \frac{2.00}{2.50} = .80$$

We can even estimate the reliability for a shorter version of the test, say a 5-item version by using a decimal fraction, that is, n = .33 (for the number of times we must multiply 15 to get 5; .33 x 15 = 4.95 or about 5), as follows:

$$reliability = \frac{2 \times r_{half-test}}{1 + r_{half-test}} = \frac{2 \times .50}{1+.50} = \frac{1.00}{1.50} = .666 \approx .67$$

We might want to use this last strategy if we were trying to figure out how short we could make our test and still maintain decent reliability. [For more on the Spearman-Brown formula, see Brown, 1996, pp. 194-196, 204-205, or Brown with Wada, 1999, pp. 220-223, 233-234.]

Conclusion

The Spearman-Brown prophecy formula can be used for adjusting split-half reliability, but more importantly, it can be used for answering what-if questions about test length when you are designing or revising a language test. Unfortunately, the Spearman-Brown formula is limited to estimating differences on one dimension (usually the number of items, or raters). For those interested in doing so on more than one dimension, generalizability theory (G-theory) provides the same sort of answers, but for more dimensions (called facets in G-theory). For instance, in Brown (1999), I used G-theory to examine (separately and together) the effects on reliability of various numbers of items and subtests on the TOEFL, and numbers of languages among the persons taking the test.

In any case, the sentence taken from Hirai (1999) about her use of the Spearman-Brown formula is just one sentence, indeed just one small detail, taken from a larger study that has much to offer readers interested in statistical research design, and/or the relationship between listening and reading rates for EFL learners in Japan. Sometimes, it is important to look past a single tree in order to see the entire forest clearly.

References

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J. D. (trans. by M. Wada). (1999). *Gengo tesuto no kisochishiki* [Basic knowledge of language testing]. Tokyo: Taishukan Shoten.

Brown, J. D. (1999). Relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16* (2), 216-237.

Brown, J. D., Cunha, M. I. A., & Frota, S. de F. N. (2001). The development and validation of a Portuguese version of the Motivated Strategies for Learning Questionnaire. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition*. Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawaii Press.

Hirai, A. (1999). The relationship between listening and reading rates of Japanese EFL learners. *Modern Language Journal, 83*, 367-384.