Shiken: JALT Testing & Evaluation SIG Newsletter. 6 (1) February 2002 (p. 17 - 18)

Statistics Corner: **Questions and answers about language testing statistics:**

The Cronbach alpha reliability estimate

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

QUESTION: For what kind of test would a coefficient alpha reliability be appropriate? How does one interpret reliability coefficients?

ANSWER: Coefficient alpha is one name for the Cronbach alpha reliability estimate. Cronbach alpha is one of the most commonly reported reliability estimates in the language testing literature. To adequately explain Cronbach alpha, I will need to address several sub-questions: (a) What are the different strategies for estimating reliability? (b) Where does Cronbach alpha fit into these strategies for estimating reliability? And, (c) how should we interpret Cronbach alpha?

What are the different strategies for estimating reliability?

As I pointed out in Brown (1997), testing books (e.g., Brown 1996, or 1999a) usually explain three strategies for estimating reliability: (a) test-retest reliability (i.e., calculating a reliability estimate by administering a test on two occasions and calculating the correlation between the two sets of scores), (b) equivalent (or parallel) forms reliability (i.e., calculating a reliability estimate by administering two forms of a test and calculating the correlation between the two sets of scores), and (c) internal consistency reliability (i.e., calculating a reliability estimate based on a single form of a test administered on a single occasion using one of the many available internal consistency equations). Clearly, the internal consistency strategy is the easiest logistically because it does not require administering the test twice or having two forms of the test.

Where does Cronbach alpha fit into these strategies for estimating reliability?

Internal consistency reliability estimates come in several flavors. The most familiar are the (a) split-half adjusted (i.e., adjusted using the Spearman-Brown prophecy formula, which is the focus of Brown, 2001), (b) Kuder-Richardson formulas 20 and 21 (also known as K-R20 and K-R21, see Kuder & Richardson, 1937), and (c) Cronbach alpha (see Cronbach, 1970).

The most frequently reported internal consistency estimates are the K-R20 and Cronbach alpha. Either one provides a sound under-estimate (that is conservative or safe estimate) of the reliability of a set of test results. However, the K-R20 can only be applied if the test items are scored dichotomously (i.e., right or wrong). Cronbach alpha can also be applied when test items are scored dichotomously, but alpha has the advantage over K-R20 of being applicable when items are weighted (as in an item scored 0 points for a functionally and grammatically incorrect answer, 1 point for a functionally incorrect, but grammatically correct answer, 2 points for a functionally correct but grammatically incorrect answer, and 3 points for a functionally and grammatically correct answer). Hence, Cronbach alpha is more flexible than K-R20 and is often the appropriate reliability estimate for language test development projects and language testing research.

How should we interpret Cronbach alpha?

A Cronbach alpha estimate (often symbolized by the lower case Greek letter α) should be interpreted just like other internal consistency estimates, that is, it estimates the proportion of variance in the test scores that can be attributed to true score variance. Put more simply, Cronbach alpha is used to estimate the proportion of variance that is systematic or consistent in a set of test scores. It can range from 00.0 (if no variance is consistent) to 1.00 (if all variance is consistent) with all values between 00.0 and 1.00 also being possible. For example, if the Cronbach alpha for a set of scores turns out to be .90, you can interpret that as meaning that the test is 90% reliable, and by extension that it is 10% unreliable (100% - 90% = 10%).

However, when interpreting Cronbach alpha, you should keep in mind at least the following five concepts:

1. Cronbach alpha provides an estimate of the internal consistency of the test, thus (a) alpha does not indicate the stability or consistency of the test over time, which would be better estimated using the test-retest reliability strategy, and (b) alpha does not indicate the stability or consistency of the test across test forms, which would be better estimated using the equivalent forms reliability strategy.
2. Cronbach alpha is appropriately applied to norm-referenced tests and norm-referenced decisions (e.g., admissions and placement decisions), but not to criterion-referenced tests and criterion-referenced decisions (e.g., diagnostic and achievement decisions).
3. All other factors held constant, tests that have normally distributed scores are more likely to have high Cronbach alpha reliability estimates than tests with positively or negatively skewed distributions, and so alpha must be interpreted in light of the particular distribution involved.
4. All other factors held constant, Cronbach alpha will be higher for longer tests than for shorter tests (as shown and explained in Brown 1998 & 2001), and so alpha must be interpreted in light of the particular test length involved.
5. The standard error of measurement (or SEM) is an additional reliability statistic calculated from the reliability estimate (as explained in Brown, 1999b) that may prove more useful than the reliability estimate itself when you are making actual decisions with test scores. The SEM's usefulness arises from the fact that it provides an estimate of how much variability in actual test score points you can expect around a particular cut-point due to unreliable variance (with 68% probability if one SEM plus or minus is used, or with 95% if two SEMs plus or minus are used, or 98% if three are used). (For more on this topic, see Brown 1996 or 1999a).

Conclusion

Clearly, Cronbach alpha is a useful and flexible tool that you can use to investigate the reliability of your language test results. In the process, it is important to remember that reliability, regardless of the strategy used to obtain it, is not a characteristic inherent in the test itself, but rather is an estimate of the consistency of a set of items when they are administered to a particular group of students at a specific time under particular conditions for a specific purpose. Extrapolating from reliability results obtained under a particular set of circumstances to other situations must be done with great care.

References

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* Cambridge: Cambridge University Press.

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

Brown, J. D. (1997). Statistics Corner: Questions and answers about language testing statistics: Reliability of surveys. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1* (2), 17-19. Retrieved December 24, 2001 from the World Wide Web: http://jalt.org/test/bro_2.htm

Brown, J. D. (1998). Statistics Corner: Questions and answers about language testing statistics: Reliability and cloze test length. *Shiken: JALT Testing & Evaluation SIG Newsletter, 2* (2), 19-22. Retrieved December 24, 2001 from the World Wide Web: http://jalt.org/test/bro_3.htm

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

Brown, J. D. (1999b). Statistics Corner. Questions and answers about language testing statistics: The standard error of vs. standard error of measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter, 3* (1), 15-19. Retrieved December 24, 2001 from the World Wide Web: http://jalt.org/test/bro_4.htm.

Brown, J. D. (2001). Statistics Corner. Questions and answers about language testing statistics: Can we use the Spearman-Brown prophecy formula to defend low reliability? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4* (3), 7-9. Retrieved December 24, 2001 from the World Wide Web: http://jalt.org/test/bro_9.htm.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika.* 2, 151-160.