Shiken: JALT Testing & Evaluation SIG Newsletter, 1 (2) September 1997 (p. 18-21)

Statistics Corner

Questions and answers about language testing statistics:

Reliability of surveys

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

QUESTION: In a *JALT Journal* article I recently read (Sasaki, 1996), the author wrote about a teacher survey. In it, she reported that she used the Cronbach alpha statistic to measure internal consistency, with a resulting alpha value of 0.70 (p. 232). When should this internal consistency measure be used and how good is a reliability figure of 0.70?

ANSWER: To fully answer your question, I need to first address several sub-questions:

* Why is consistency important?
* Does the notion of consistency vary in different situations?
* What do norm-referenced reliability statistics mean?
* What factors affect the reliability of an instrument?
* And what kinds of norm-referenced reliability statistics are appropriate for questionnaires?

Before I am finished here, I do promise to come back to your original question.

Why is consistency important?

For language surveys, like any other measurement instrument, we need to be concerned about the consistency of measurement. This concern should apply to all observations or measurements that we make, whether they be qualitative or quantitative, and whether they be for pedagogical/ curriculum purposes or for research purposes.

In a sense, we are no different from a post office that weighs letters and packages. The customers at a post office quite reasonably expect the scales to be consistent. If they take a package to the post office and weigh it on two different occasions, they expect the two weights to be the same, or at least very similar. Thus measurement consistency is an important issue in everyday life. If fact, measurement consistency is such an important issue that national governments have established departments of weights and measures to insure measurement consistency. Similarly, we would like our measurements in language teaching to be consistent.

Does the notion of consistency vary in different situations?

Measurement consistency may be defined differently depending on the situation involved. Qualitative researchers refer to the consistency concept as dependability and will stress the importance of using multiple sources of information, especially triangulation (wherein different types of information are gathered to cross validate each other; for instance, three different kinds of information, say interviews, observations, and questionnaires, might be gathered from three different sources, say students, teachers, and administrators).

Quantitative researchers also refer to the consistency concept as dependability for criterion-referenced assessments, but they call it reliability for norm-referenced assessments. Criterion-referenced dependability typically involves one or more of the following three strategies (these three will not be defined here because they are not directly germane to the discussion at hand): threshold-loss agreement, squared-error loss, or domain-score dependability. Norm-referenced reliability usually involves one or more of following three strategies: test-retest reliability, equivalent forms reliability, or internal consistency reliability (these three strategies will be defined below). Like qualitative researchers, quantitative researchers should stress the importance of multiple sources of information, especially in making important decisions about students' lives. (For much more information, including how to calculate the various reliability and dependability estimates, see Brown, 1996).

Clearly, for any observations or measurements (whether qualitative or quantitative), consistency is a crucial concept for language teachers, administrators, and researchers of all kinds. Equally clear is the fact that the concept of consistency is approached differently depending on the type of measurement involved and the purposes of that measurement.

What do norm-referenced reliability statistics mean?

Zeroing in on the meaning of reliability statistics for norm-referenced purposes, they are most commonly reported as the proportion of consistent variance on whatever instrument is being examined. These reliability statistics range from zero (for a totally unreliable measure) to 1.00 (for a measure that is 100% reliable). Thus, an estimate of .70, like the one you asked about, indicates that the questionnaire is 70% reliable. Note that a .70 estimate also means that the measure is 30% unreliable (1 - .70 = .30, or 30%). The question of whether such a value is good or bad can only be answered in relative terms depending on a number of factors.

What factors affect the reliability of an instrument?

Reliability is affected by many factors, but from the researcher's point of view, the three most important factors are the length (or total number of questions), the quality of the questions, and the fit to the group being measured. If a test or questionnaire is reasonably long (say at least 30 questions) and is well-written, it should be fairly reliable, but if the instrument is short and/or the questions are not effective, it will not be reliable at all.

Furthermore, even if sufficient well-written questions are used, an instrument may prove to be unreliable if it does not fit the group of people involved in the measurement process. So a test like the TOEFL that has repeatedly been shown to be highly reliable (in excess of .90 when administered to students ranging from near zero English to native-like proficiency) may be much less reliable if used in a local English language program where it doesn't fit very well, that is, where the range of abilities is very restricted or where the scores are skewed. [Restricted range means the scores don't vary much. Skewed, in non-technical terms, means that the scores are scrunched up toward the top or bottom of the score range. [For more complete definitions, see Brown, 1996.] We

will return to these issues, as they apply to the .70 reported in Sasaki's (1996) study, at the end of this article.

What kinds of norm-referenced reliability statistics are appropriate for questionnaires?

Testing books (e.g., Brown, 1996) typically discuss test-retest reliability (where the researcher administers a measure on two occasions and calculates the correlation between the two sets of scores as a reliability estimate), equivalent forms reliability (where the researcher administers two forms of a measure and calculates the correlation between the two sets of scores as a reliability estimate), and internal consistency reliability (where the researcher estimates the reliability of a single form administered on a single occasion). Obviously the internal consistency estimates are the easiest to get because it is not necessary to administer the measure twice or to have two forms.

Internal consistency reliability estimates come in many forms, e.g., the split-half adjusted (using the Spearman-Brown prophecy formula), the Kuder-Richardson formulas 20 and 21 (aka, K-R20 and K-R21), and Cronbach alpha. The most commonly reported of these are the K-R20 and the Cronbach alpha. Either one provides a sound estimate of reliability. However, the K-R20 is applicable only when questions are scored in a binary manner (i.e., right or wrong). Cronbach alpha has the advantage of being applicable when questions are small scales in their own right like the Likert scale (i.e., 1 2 3 4 5 type) questions found on many questionnaires. Hence, Cronbach alpha is most often the reliability estimate of choice for survey research.

A direct answer to your question

Coming back to your question, the central parts were "Why would someone want to use a measure of internal consistency for survey data, and is 0.70 good or bad?"

Why measure internal consistency for survey data? A direct answer to this question is that the researcher probably wanted to know about the consistency of her survey instrument because the results of her study rested entirely on that measurement. If the consistency of her instrument was questionable, then all of her results would be equally questionable. Put another way, her results could be no more reliable than the instrument upon which they were based.

Is .70 good or bad? The answer to this part of your question is "it depends." Ultimately, you, the reader of such an article, must decide whether you think .70 is good or bad. Is 70% reliability good? That will depend on the length of the questionnaire, the quality of the questions, and their fit to the group being measured. For what they are worth, my reactions as a reader of Sasaki's article are as follows:

In terms of length, there are 25 questions, which would lead me to think that the reliability might be higher if she had just a few more questions.

Looking at question quality, I note that the questionnaire was developed solely out of the researcher's experience and was a first attempt, so I would not expect it to be extremely reliable. In her favor, she does present the actual questions (in her Tables 1 and 2), so readers can directly inspect them. And, the questions strike me as being reasonably well-crafted.

In addition, considering the fit to the group, the questionnaire appears to fit the group in the sense that it produced considerable systematic variance as indicated by the fact that the mean answers on the various five-point Likert-scale questions ranged from 1.48 to 4.83 with many different values in between.

Finally, the results of the questionnaire are systematic enough to have produced significant differences between preferred and perceived student behaviors on 24 out of the 25 questions. So the fit of the questionnaire appears to be fairly good.

But, that is just one man's opinion. What do you think? Whatever you decide will (and should) affect the way you read and interpret the rest of the study. So this is a rather critical decision. Maybe the question shouldn't be so much whether .70 is good or bad (in absolute terms), but rather whether 70% reliability is sufficient for the purposes of the study involved. From that perspective, I would have to say that 70% reliability is not spectacular, but it does seem sufficiently high for the purposes of Sasaki's survey study.

In any case, my hat is off to her for reporting the reliability of her instrument. Such statistics are seldom reported in language teaching journals. Thus we often have no idea how good the instruments (upon which entire studies are based) really are in terms of reliability. I also congratulate the author for showing her questions directly to the readers. There is no better way to decide for yourself whether a questionnaire makes sense to you and to your teaching situation, than to inspect the questions directly. By reporting her reliability and including her questions, Sasaki was being open and honest about her study and giving her readers sufficient information to judge its quality. Bravo!

References

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: Skewness and kurtosis. *Shiken 1* (1), 20-23. Available online at www.jalt.org/test/bro_1.htm. [16 Aug. 1997].

Sasaki, C. L. (1996). Teacher preferences of student behavior in Japan. *JALT Journal, 18* (2), 229-239.