Shiken: JALT Testing & Evaluation SIG Newsletter, 4 (2) Oct 2000 (p. 8 - 12)

Statistics Corner

Questions and answers about language testing statistics:

What is construct validity?

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

QUESTION: Recently I came across an article mentioning that a test had poor construct validity. What exactly is construct validity? How well accepted is the concept of construct validity? How does it differ from other forms of validity? What is the best way of measuring construct validity? And finally, what are the most common threats to construct validity?

ANSWER: The general concept of validity was traditionally defined as "the degree to which a test measures what it claims, or purports, to be measuring" (Brown, 1996, p. 231). However, as your questions indicate, the issues involved in validity are not that simple. To address these issues head on, I will use your questions as headings and take the liberty of rearranging them a bit.

How does construct validity differ from other forms of validity?

Validity was traditionally subdivided into three categories: content, criterion-related, and construct validity (see Brown 1996, pp. 231-249). *Content validity* includes any validity strategies that focus on the content of the test. To demonstrate content validity, testers investigate the degree to which a test is a representative sample of the content of whatever objectives or specifications the test was originally designed to measure. To investigate the degree of match, test developers often enlist well-trained colleagues to make judgments about the degree to which the test items matched the test objectives or specifications.

Criterion-related validity usually includes any validity strategies that focus on the correlation of the test being validated with some well-respected outside measure(s) of the same objectives or specifications. For instance, if a group of testers were trying to develop a test for business English to be administered primarily in Japan and Korea, they might decide to administer their new test and the TOEIC® to a fairly large group of students and then calculate the degree of correlation between the two tests. If the correlation coefficient between the new test and the TOEIC turned out to be high, that would indicate that the new test was arranging the students along a continuum of proficiency levels very much like the TOEIC does – a result that could, in turn, be used to support the validity of the new test. Criterion-related validity of this sort is sometimes called concurrent validity (because both tests are administered at about the same time).

Another version of criterion-related validity is called *predictive validity*. Predictive validity is the degree of correlation between the scores on a test and some other measure that the test is designed to predict. For example, a number of studies have been conducted to examine the degree of relationship between students' Graduate Record Examination® (GRE) scores and their grade point averages (GPA) after two years of graduate study. The correlation between these two variables represents the degree to which the GRE predicts academic achievement as measured by two years of GPA in graduate school.

What exactly is construct validity?

To understand the traditional definition of construct validity, it is first necessary to understand what a construct is. A construct, or psychological construct as it is also called, is an attribute, proficiency, ability, or skill that happens in the human brain and is defined by established theories. For example, "overall English language proficiency" is a construct. It exists in theory and has been observed to exist in practice.

Construct validity has traditionally been defined as the experimental demonstration that a test is measuring the construct it claims to be measuring. Such an experiment could take the form of a differential-groups study, wherein the performances on the test are compared for two groups: one that has the construct and one that does not have the construct. If the group with the construct performs better than the group without the construct, that result is said to provide evidence of the construct validity of the test. An alternative strategy is called an intervention study, wherein a group that is weak in the construct is measured using the test, then taught the construct, and measured again. If a non-trivial difference is found between the pretest and posttest, that difference can be said to support the construct validity of the test. Numerous other strategies can be used to study the construct validity of a test, but more about that later.

"[The] unified view of construct validity is considered a new development by many of the language testers around the world. However, it can hardly be new . . . "

How well accepted is the concept of construct validity?

The concept of construct validity is very well accepted. Indeed, in educational measurement circles, all three types of validity discussed above (content, criterion-related, and construct validity) are now taken to be different facets of a single unified form of construct validity. This unified view of construct validity is considered a new development by many of the language testers around the world. However, it can hardly be new given that I remember discussing it in courses I took with Richard Shavelson at UCLA in the late 1970s.

Coming back to your question, either the traditional view of construct validity or the unified view is held by virtually all psychometricians inside or outside of language testing. Thus, construct validity can be said to be well-accepted, one way or the other.

What is the best way of measuring construct validity?

Regardless of how construct validity is defined, there is no single best way to study it. In most cases, construct validity should be demonstrated from a number of perspectives. Hence, the more strategies used to demonstrate the validity of a test, the more confidence test users have in the construct validity of that test, but only if the evidence provided by those strategies is convincing.

In short, the construct validity of a test should be demonstrated by an accumulation of evidence. For example, taking the unified definition of construct validity, we could demonstrate it using content analysis, correlation coefficients, factor analysis, ANOVA studies demonstrating differences between differential groups or pretest-posttest intervention studies, factor analysis, multi-trait/multi-method studies, etc. Naturally, doing all of the above would be a tremendous amount of work, so the amount of work a group of test developers is willing to put into demonstrating the construct validity of their test is directly related to the number of such demonstrations they can provide. Smart test developers will stop when they feel they have provided a convincing set of validity arguments.

What are the most common threats to construct validity?

Any threats to the reliability (or consistency) of a test are also threats to its validity because a test cannot be said to be any more systematically valid than it is first systematic (or consistent). Thirty-six such threats to reliability are discussed in detail in Brown (1996, pp. 188-192) in five different categories of problems due to the: environment of the test administration, administration procedures, examinees, scoring procedures, and test construction (or quality of test items).

In my view, the validity problems I have most often observed in Japan are an inadequate number of items, poor item writing, lack of pilot testing, lack of item analysis procedures, lack of reliability studies, and lack of validity analysis. These are all problems that could be rectified by using the well-developed psychometric procedures used in many countries around the world.

Conclusion

". . . the validity problems I have most often observed in Japan are an inadequate number of items, poor item writing, lack of pilot testing, lack of item analysis procedures, lack of reliability studies, and lack of validity analysis. These are all problems that could be rectified by using the well-developed psychometric procedures used in many countries around the world."

In discussing language test validity at this point in time, I would be remiss to not at least mention Messick's (1988, 1989) thinking about validity. Messick presented a unified and expanded theory of validity, which included the evidential and consequential bases of test interpretation and use. Table 1 shows how this theory works. Notice that the evidential basis for validity includes both test score interpretation and test score use. The evidential basis for interpreting tests involves the empirical study of construct validity, which is defined by Messick as the theoretical context of implied relationships to other constructs. The evidential basis for using tests involves the empirical investigation of both construct validity and relevance/utility, which are defined as the theoretical contexts of implied applicability and usefulness.

Table 1. *Facets of test validity according to Messick*. (Adapted from ******)

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance and Utility
Consequential Basis	Value Implications	Social Consequences

The consequential basis of validity involves both test score interpretation and test score use. The consequential basis for interpreting tests requires making judgments of the value implications, which are defined as the contexts of implied relationships to good/bad, desirable/undesirable, etc. score interpretations. The consequential basis for using tests involves making judgments of social consequences, which are defined as the value contexts of implied consequences of test use and the tangible effects of actually applying that test. The value implications and social consequences issues have special importance in Japan, where the values underlying tests like the university entrance exams and the social consequences of their use are so omnipresent in educators minds. (For more information on this model of validity, see Messick, 1988, 1989; for some interesting discussions of the consequential aspects of validity, see Green, 1998; Linn, 1998; Lune, Parke, & Stone, 1998;

Moss, 1998; Reckase, 1998; Taleporos, 1998; and Yen, 1998.)

Clearly then, while construct validity is still an important concept, our responsibilities as language testers appear to have expanded considerably with Messick's call for test developers to pay attention to the evidential and consequential bases for the use and interpretation of test scores.

References

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

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement, 17* (2),16-19.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement, 17* (2), 28-30

Lune, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement, 17* (2), 24-28.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement, 17*(2), 6-12.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement, 17* (2), 13-16.

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement, 17* (2), 20-23.

Yen, W. M. (1998). Investigating the consequential aspects of validity: Who is responsible and what should they do? *Educational Measurement, 17* (2), 5-6.

HTML: http://www.jalt.org/test/bro_8.htm **PDF**: http://www.jalt.org/test/PDF/Brown8.pdf