Language tests reliability and validity

A test is reliable to the extent that whatever it measures, it measures it consistently. If I were to stand on a scale and the scale read 15 pounds, I might wonder. Suppose I were to step off the scale and stand on it again, and again it read 15 pounds. The scale is producing consistent results. From a research point of view, the scale seems to be reliable because whatever it is measuring, it is measuring it consistently. Whether those consistent results are valid is another question. However, an instrument cannot be valid if it is not reliable.

The real difference between reliability and validity is mostly a matter of definition. Reliability estimates the consistency of your measurement, or more simply the degree to which an instrument measures the same way each time it is used in under the same conditions with the same subjects. Validity, on the other hand, involves the degree to which your are measuring what you are supposed to, more simply, the accuracy of your measurement. It is my belief that validity is more important than reliability because if an instrument does not accurately measure what it is supposed to, there is no reason to use it even if it measures consistently (reliably).

There are three major categories of reliability for most instruments: test-retest, equivalent form, and internal consistency. Each measures consistency a bit differently and a given instrument need not meet the requirements of each. Test-retest measures consistency from one time to the next. Equivalent-form measures consistency between two versions of an instrument. Internal-consistency measures consistency within the instrument (consistency among the questions). A fourth category (scorer agreement) is often used with performance and product assessments. Scorer agreement is

consistency of rating a performance or product among different judges who are rating the performance or product. Generally speaking, the longer a test is, the more reliable it tends to be (up to a point). For research purposes, a minimum reliability of . 70 is required. Some researchers feel that it should be higher. A reliability of 0. 70 indicates 70% consistency in the scores that are produced by the instrument. Many tests, such as achievement tests, strive for 0. 90 or higher reliabilities.

Relationship of Test Forms and Testing Sessions Required for Reliability Procedures

Testing Sessions Required

Test Forms Required

One

Two

One

Two

Split-Half

Kuder-Richardson

Cronbach’s Alpha

Equivalent (Alternative) Form

Test-Retest

1 Test-Retest Method. The same instrument is given twice to the same group of people. The reliability is the correlation between the scores on the two instruments. If the results are consistent over time, the scores should be similar. The trick with test-retest reliability is determining how long to wait between the two administrations. One should wait long enough so the subjects don’t remember how they responded the first time they completed the instrument, but not so long that their knowledge of the material being measured has changed. This may be a couple weeks to a couple months.

2 Equivalent-Form (Parallel or Alternate-Form) Method. Two different versions of the instrument are created. We assume both measure the same thing. The same subjects complete both instruments during the same time period. The scores on the two instruments are correlated to calculate the consistency between the two forms of the instrument.

3 Internal-Consistency Method. Several internal-consistency methods exist. They have one thing in common. The subjects complete one instrument one time. For this reason, this is the easiest form of reliability to investigate. This method measures consistency within the instrument three different ways.

– Split-Half. A total score for the odd number questions is correlated with a total score for the even number questions (although it might be the first half with the second half). This is often used with dichotomous variables that are scored 0 for incorrect and 1 for correct. The Spearman-Brown prophecy formula is applied to the correlation to determine the reliability.

– Kuder-Richardson Formula 20 (K-R 20) and Kuder-Richardson Formula 21 (K-R 21). These are alternative formulas for calculating how consistent

subject responses are among the questions on an instrument. Items on the instrument must be dichotomously scored (0 for incorrect and 1 for correct). All items are compared with each other, rather than half of the items with the other half of the items. It can be shown mathematically that the Kuder-Richardson reliability coefficient is actually the mean of all split-half coefficients resulting from different splittings of a test. K-R 21 assumes that all of the questions are equally difficult. K-R 20 does not assume that.

– Cronbach’s Alpha, also known as Coefficient Alpha. When the items on an instrument are not scored right versus wrong, Cronbach’s alpha is often used to measure the internal consistency. This is often the case with attitude instruments that use the Likert scale. A computer program such as SPSS is often used to calculate Cronbach’s alpha. Although Cronbach’s alpha is usually used for scores which fall along a continuum, it will produce the same results as KR-20 with dichotomous data (0 or 1).

4 Scorer Agreement. Performance and product assessments are often based on scores by individuals who are trained to evaluate the performance or product. The consistency between rating can be calculated in a variety of ways.

– Interrater Reliability. Two judges can evaluate a group of student products and the correlation between their ratings can be calculated (r=. 90 is a common cutoff).

– Percentage Agreement. Two judges can evaluate a group of products and a percentage for the number of times they agree is calculated (80% is a common cutoff).

All scores contain error. The error is what lowers an instrument's reliability: Obtained (also " observed") Score = True Score + Error Score. There could be a number of reasons why the reliability estimate for a measure is low. Four common sources of inconsistencies of test scores are distinguished: (a) test taker – perhaps the subject is having a bad day, (b) test Itself – the questions on the instrument may be unclear, (c) testing conditions – there may be distractions during the testing that detract the subject, (d) test scoring – scorers may be applying different standards when evaluating subjects' responses.

Reliability

Validity

An instrument is valid only to the extent that it's scores permit appropriate inferences to be made about (a) a specific group of people for (b) specific purposes.

An instrument that is a valid measure of third grader's language skills probably is not a valid measure of high school student's language proficiency. An instrument that is a valid predictor of how well students might do in school, may not be a valid measure of how well they will do once they complete school. So we never say that an instrument is valid or not valid. We say it is valid for a specific purpose with a specific group of people. Validity is specific to the appropriateness of the interpretations we wish to make with the scores. For example, a measuring tape is a valid instrument to determine people's height; it is not a valid instrument to determine their weight.

There are three general categories of instrument validity.

1 Content-Related Evidence (also known as Face Validity). Specialists in the content measured by the instrument are asked to judge the appropriateness of the items on the instrument. Do they cover the breath of the content area (does the instrument contain a representative sample of the content being assessed)? Are they in a format that is appropriate for those using the instrument? A test that is intended to measure the quality of science instruction in fifth grade, should cover material covered in the fifth grade science course in a manner appropriate for fifth graders. A national science test might not be a valid measure of local science instruction, although it might be a valid measure of national science standards.

2 Criterion-Related Evidence. Criterion-related evidence is collected by comparing the instrument with some future or current criteria, thus the name criterion-related. The purpose of an instrument dictates whether predictive or concurrent validity is warranted.

– Predictive Validity. If an instrument is purported to measure some future performance, predictive validity should be investigated. A comparison must be made between the instrument and some later behavior that it predicts. Suppose a screening test for 5-year-olds is purported to predict success in kindergarten. To investigate predictive validity, one would give the prescreening instrument to 5-year-olds prior to their entry into kindergarten. The children’s kindergarten performance would be assessed at the end of kindergarten and a correlation would be calculated between the screening instrument scores and the kindergarten performance scores.

– Concurrent Validity. Concurrent validity compares scores on an instrument with current performance on some other measure. Unlike predictive validity, where the second measurement occurs later, concurrent validity requires a second measure at about the same time. Concurrent validity for a science test could be investigated by correlating scores for the test with scores from another established science test taken about the same time. Another way is to administer the instrument to two groups who are known to differ on the trait being measured by the instrument. One would have support for concurrent validity if the scores for the two groups were very different. An instrument that measures altruism should be able to discriminate those who possess it (nuns) from those who don't (homicidal maniacs). One would expect the nuns to score significantly higher on the instrument.

3 Construct-Related Evidence. Construct validity is an on-going process. The possible extremes are:

– Discriminant Validity. An instrument does not correlate significantly with variables from which it should differ.

– Convergent Validity. An instrument correlates highly with other variables with which it should theoretically correlate.

Note that recent research has shown the unitary nature of the construct of validity.