Reliability in Language Assessment

Carol Chapelle
Iowa State University, carolc@iastate.edu

Follow this and additional works at: http://lib.dr.iastate.edu/engl_pubs

Part of the Language and Literacy Education Commons

The complete bibliographic information for this item can be found at http://lib.dr.iastate.edu/engl_pubs/207. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.
Reliability in Language Assessment

**Abstract**
Almost any test user will readily agree that tests should be valid and reliable, but if asked to explain what "reliability" means, most test users would be hard pressed, despite the widespread use of this measure in the documentation that accompanies language tests. Assessment experts would also agree that reliability is a central concern for interpreting assessment results, even to the point that it is an important part of most validity arguments. However, the experts would also point out that reliability is not a characteristic of a test, but rather a characteristic of test scores obtained from a given test administration or administrations.

**Disciplines**
Language and Literacy Education

**Comments**
Almost any test user will readily agree that tests should be valid and reliable, but if asked to explain what “reliability” means, most test users would be hard pressed, despite the widespread use of this measure in the documentation that accompanies language tests. Assessment experts would also agree that reliability is a central concern for interpreting assessment results, even to the point that it is an important part of most validity arguments. However, the experts would also point out that reliability is not a characteristic of a test, but rather a characteristic of test scores obtained from a given test administration or administrations.

Reliability, which means consistency or stability in language assessment, is most simply explained with reference to classical test theory, which posits a true score that an examinee would obtain on a test that was a perfectly reliable measure of an attribute. Such reliable tests exist only in theory but, hypothetically, if the attribute of interest were assessed by...
gathering many samples of performance on many tests of the attribute, a close approximation of the true score could be made. As Haertel explains,

Any of these possible scores would have served the purpose of measurement equally well, but they would not all be identical. Taken together, this hypothetical collection of scores represents the general, enduring attribute of interest. Thus, it is important to determine the extent to which any single result of the measurement procedure is likely to depart from the average score over many replications. Other things being equal, the greater uncertainty associated with the results of the measurement, the less confidence should be placed in that measurement. (Haertel, 2006, p. 65)

The assumption underlying this perspective is that sources of inconsistency in test scores produce measurement error in the test scores. This error can come from factors such as difficult conditions for listening during a listening test, fatigue on the part of examinees, unclear instructions in one section of a test, poorly written test tasks, inadequate scoring rubrics on a writing test, or simply an inadequate sample of performance. The assumption is that any of these conditions arising from test development, test administration, and test scoring will affect scores randomly whereas performance based on the defined attribute (e.g., language ability, listening comprehension, vocabulary knowledge) will affect scores consistently. Estimating reliability, then, becomes a matter of sorting out the amount of consistency in scores from the amount of error.

Estimating Reliability

Methods for estimating reliability within a classical test theory perspective depend on the sources of error that one wishes to isolate, but each method is based on the fundamental idea of replication that Haertel described. Brennan (2001) argues that “the notion of replications is central to an understanding of reliability” because “reliability is a measure of the degree of consistency in examinee scores over replications of a measurement procedure” (pp. 295–6). In order to understand the reliability of any particular measure, then, the test developer needs to be able to define what a replication of the test in question would consist of. The clearest type of replication in testing is two administrations of the same test. A reliability coefficient is calculated as the correlation between the two sets of scores obtained by examinees who took the test twice, and such a reliability estimate is called the test–retest reliability. Parallel-forms methods require two forms of a test to be administered to the same examinees so the consistency of performance across forms can be estimated. In both cases the statistical procedure for estimating reliability is a correlation between the two sets of scores. Despite the methodological clarity of test–retest (and parallel-forms) reliability, in operational testing programs it is difficult to obtain two sets of test scores on exactly the same test or a parallel one from the same group of examinees. Instead, a variation of a split-half method is used in many operational settings.

Split-half methods treat a half of a test as a replication of the other half, and are therefore also called internal consistency reliability methods. Cronbach’s alpha is the most widely used of the many methods for estimating internal consistency reliability in language testing and elsewhere. It can be calculated through the use of SPSS, for example, when the researcher has the item scores (dichotomous or polytomous) for each test taker. Inter-rater methods assess the consistency between the ratings given by two or more raters to a set of test performances, typically using a correlation. Observed reliabilities are affected by the manner in which the reliability was estimated. When the two testing events are separated by time, as in test–retest coefficients, they tend to be lower than when all samples of performance are obtained in a single testing event (Gronlund, 1985).
In each of these cases a statistic estimating the degree of relationship, or consistency across the repetitions, is used to indicate the reliability of the test scores. It is expressed as a value that can range from 0, meaning variation in test scores is completely random, to 1, indicating that variation is completely reliable, that is, attributable to one factor. Reliabilities of .90 and higher are considered very good, whereas reliabilities of below .70 indicate scores containing a lot of error. Whether or not the amount of error is too great, of course, depends on the nature of the test and its use. For a high-stakes test such as the Internet-based Test of English as a Foreign Language (TOEFL iBT®), which is used for admission decisions, an internal consistency reliability of .94 is reported for total scores (Enright & Tyson, 2011). However, this test takes hours to administer and contains multiple sections with multiple test tasks. Shorter tests with fewer items typically do not produce scores with such high reliability estimates. Such tests may still be entirely appropriate and useful for their intended purposes.

Applied linguists can find a discussion of reliability and instructions for calculating reliability in any current introductions to language testing, but Bachman (2004) and Brown (2005) provide the most thorough treatment. However, applied linguists tend to be interested not only in calculating reliability but also in interpreting it. They are therefore interested in understanding how observation of consistent performance can accurately reflect an attribute such as language ability that has so many inherent inconsistencies.

### Should Language Performance Be Consistent?

Books on language assessment that present an in-depth perspective on the construct of language ability (Bachman, 1990; McNamara, 1996) hint at a potential contradiction between the way the construct is defined and the expectation that a good measure will exhibit consistencies in performance. Bachman’s presentation of the multifaceted communicative language ability posits multiple factors including knowledge of aspects of language (e.g., vocabulary, pragmatics), strategic competence, and topical knowledge as part of the construct. If the construct of communicative language ability is itself so multifaceted and variable across language use contexts, why do language testers hold reliability up as an ideal quality in test performance?

In recognition of this apparent contradiction, Swain (1993) questioned the relevance of internal consistency reliability for test scores which are intended to reflect examinees’ communicative language proficiency. If the construct of communicative language proficiency would not itself be expected to be internally consistent because it consists of components that do not necessarily develop simultaneously, how can consistency in test performance across tasks, occasions, or forms be sought? The consistency of measurement versus variability of construct meaning is a theme that is taken up regularly in applied linguistics, where many teachers and researchers see the variability of language ability as more evident and pertinent than the consistency, and thus question the utility of language tests, whose raison d’être is to provide scores that offer a consistent summary of examinees’ language ability.

One approach to reliability taken by Bachman and Palmer (1996) is to recognize that the way the construct is defined will affect the expected level of reliability in scores. “If the construct definition focuses narrowly on components of language ability, the test developer can reasonably expect to achieve higher levels of reliability than if the construct is complex, including a wide range of components of language ability, as well as topical knowledge” (p. 135). Therefore in determining an acceptable degree of unreliability for a test, rather than categorically seeking maximum consistency, test developers are to analyze their tests to make sure that they do not produce scores reflecting unmotivated inconsistency. The
idea that reliability needs to be evaluated in view of the construct definition and its degree of expected variability is useful in both practice and research.

Another useful perspective comes from Snow and Lohman (1989), who distinguished between the type of psychological model containing many components and processes that underlie performance and the psychometric model that is constructed to capture response consistencies. Mislevy (2006) takes this farther by displaying graphically the different levels of models that applied linguists may work with, from the delicately defined linguistic knowledge and processes, to a rougher construct description representing some of the primary components, to a psychometric model which can be expressed mathematically. Mislevy's approach emphasizes that assessments use models in order to capture a view of the attribute that is useful for some purpose. In this sense all models are inaccurate because they fail to capture some of the detail of the attribute; however, the criterion for the success of the model is not its correctness, but its usefulness. What is needed is models that are not so wrong as to not be useful, and for most purposes models underlying tests need to be consistent.

In contrast to these construct-relevance perspectives to reliability is another argument about the exclusive definition of reliability as replication. Moss (1994) notes that for many types of high-stakes decisions such as selecting manuscripts to appear in journals and applicants for academic jobs, the manner in which quality decisions are made does not draw upon the assumption that evaluators serve as replications. Instead, she argues that quality ratings are obtained by engaging in a hermeneutic process of expert judgment where experts apply their own judgments to arrive at an evaluation which may or may not agree with that of the whole group. Moss questions the unconditional value placed on reliability-as-replication as an indicator of test quality within validity arguments. She would rather “consider it one alternative for serving important epistemological and ethical purposes—an alternative that should always be justified in critical dialogue and in confrontation with other possible means of warranting knowledge claims” (Moss, 1994, p. 10).

### Using Reliability Estimates

Reliability estimates are used regularly in the everyday business of language testing and in research that uses language test scores as one source of data. These uses of reliability attest to its utility in helping test users interpret individual test scores and the performance of a set of test scores in research. Reliability is fundamental to understanding the precision of individual test scores as indicators of an attribute because it is used to calculate the standard error of measurement, a metric directly interpretable as a range of test scores relative to the observed score where the examinee’s true score probably lies. For example, the reliability of the TOEFL iBT noted above is placed into a formula to return a value of 5.64 as the standard error of measurement. This means that an individual who obtains a score on the TOEFL iBT can be assumed to have a true score within $+5.64$ to $-5.64$ of that score, with 68% probability. If the reliability of the TOEFL iBT scores were lower (i.e., less precise) the standard error of measurement would be greater, meaning that for any observed score, the true score would be within a larger range. The standard error of measurement is also useful for reminding test users that examinees’ test scores (i.e., their observed scores) are only an approximation of the scores that the examinees would obtain based on their abilities if they were to take the test repeatedly so as to obtain a more accurate estimate of their ability. This theoretical score is called the true score, and it is something that can only be estimated, not known.

In research, reliability provides an indication of the extent to which scores obtained on a measure actually reflect one construct, in other words the extent of the true score ability
that is the source of variation in scores. This is important because if a researcher is claiming to use a test score to mean one construct, such as grammatical knowledge, speaking ability, or communicative competence, and the score is intended to distinguish among test takers on the basis of their level of one construct, reliability evidence is needed. Moreover, statistically, the reliability of the measures in research affects the performance of the scores in other statistical procedures in correlational and experimental research. Both types of research seek to find systematic differences among test takers, but cannot do so unless the measures produce scores that distinguish consistently among test takers.

Because reliable test scores are most useful, one of the chief goals of test development and administration is to maximize reliability by assuring systematicity in all aspects of test development and administration. The chief rule of thumb is that the more samples of performance (i.e., the more test tasks) the more reliable the test scores. This rule follows from the basic definition of reliability as consistency across repetitions. Second, test developers attempt to minimize error associated with confusion about test tasks by trialing test items using just a few test takers to make sure tasks are clear. They also pretest items on a sample that is similar to the intended test takers because it is the set of test scores for which reliability is estimated rather than the test itself. This means that a test that produces reliable scores for one group of students may not for a second group with different characteristics.

It is worth repeating that reliability is not a characteristic of a test, but a characteristic of a set of scores. Test developers also conduct analyses of the items to identify those that do not contribute to the reliability of the measure.

**Limitations of the Classical Test Theory Perspectives**

The fundamental concepts about consistency of test scores can be expressed from a classical test theory perspective, but today other approaches are used as well. Item response theory (Lord, 1980) provides item-level information about reliability that is useful for test development, particularly development of computer-adaptive testing. Generalizability theory offers a means of isolating within one model components of error associated with different sources (Bachman, 2004). Structural equation modeling (Kline, 2005) tests the reliability of a measurement model within a more complex structural model. In all of these approaches, the language test is intended to capture what is consistent in the variation among a group for norm-referenced interpretations.

In classroom tests, such as a final exam, however, tests are intended for criterion-referenced interpretations and therefore the assumption that the desirable set of test scores should spread examinees across a normal distribution does not hold. In classroom tests, test takers' performance should be evaluated against a criterion of how much knowledge, skill, or understanding they displayed on the test. There remains an issue of consistency, however, because one would want examinees' scores to be consistent if they took the test twice or if different questions were used to elicit the performance concerning a particular ability. When criterion-referenced tests are used, consistency is estimated through the use of a metric of dependability (Brown & Hudson, 2002).

Despite the variety of approaches taken to reliability in different situations today, the basic concepts about reliability from a classical test theory perspective remain important for understanding the fundamental idea of what consistency means and why it is important.

SEE ALSO: Correlational Research in Language Assessment; Generalizability Theory in Language Testing; Modeling Language for Assessment; Statistical Analysis of Test Results; Structural Equation Modeling in Language Assessment; Validation of Language Assessments
References


Suggested Readings


Religion

SOHAIL KARMANI

Of all the manifestations of human culture, language and religion are by far two of the most striking and most enduring—so much so that they are likely to feature very prominently as markers of group cultural identity, and in so doing inform our basic ideas as to