20

The Concept of Reliability in Language Testing: Issues and Solutions

By

ROSEMARY E. CHIEDU
Department of Languages,
Delta State Polytechnic,
Ogwashi-Uku.

And

HAPPY D. OMENOGOR
Department of English,
College of Education,
Agbor,
Delta State.

Abstract

Reliability is concerned with how we measure. The reliability of a test is concerned with the consistency of scoring and the accuracy of the administration procedures of the test. In this paper, the following aspects are dealt with; the definition of language testing, types of language tests based on specific purpose and orientation, the concept of reliability in language testing, factors affecting the reliability of language tests and ways of improving reliability in language testing. All these are geared towards improving teaching and learning of students in English language classes.

What is Language Testing?

A test is a sample of behaviour. Therefore, a language test would be a sample of language behaviour. According to Hornberger and Shohamy (2008), language testing or language assessment is a field of study under the umbrella of Applied Linguistics. Its main focus is the assessment of first, second or other languages in the school, college or university context, assessment of language use in the workplace, and assessment of language in immigration, citizenship, and asylum contexts.

Language testing is the practice of evaluating the proficiency of an individual in using a particular language effectively. Language tests work best when they are designed and developed to measure specific language skills such as speaking, listening, writing proficiency, reading comprehension, the ability to translate texts, or the ability to interpret spoken language.

Theoretical Framework

The theory of language propounded by Chomsky (1963) known as Generative Grammar and Dell Hymes' Communicative Competence theory form the theoretical framework of the study. The notions of competence and performance are essential in understanding language testing or assessment. For Chomsky, competence is the capacity to generate an infinite number of sentences from a limited set of grammatical rules. This view posits that competence is logically prior to performance and is therefore, the generative basis for further learning. Dell Hymes reacted against the perceived inadequacy of Noam Chomsky's theory by stating that the communicative function of language supercedes the linguistic function of language. Communicative competence, hence, refers to a language user's knowledge of syntax, morphology, phonology, as well as social knowledge about how and when to use utterances appropriately. Language tests are conducted to test students' linguistic and communicative competence to enable them function properly in the school, work place and any other situation where the use of English is required.

Kinds of Language Tests

I. Based on specific purposes, language tests are grouped into the following types:
Proficiency test
Achievement test
Placement test
Diagnostic test
Language Aptitude test

Proficiency tests are used to get a general picture of a student's knowledge and ability rather than measure progress made in the language study. For instance, learners may be tested to find out whether their general command of English is adequate for a successful course of study in a foreign university where the language of instruction is basically English, get a job or obtain some kind of certificate. Test of English as a Foreign Language (TOEFL) is a good example of a proficiency test. In proficiency tests, the question asked is whether the knowledge of English possessed by students is good enough to cope with foreseen demands and a way to try and assess this is by representative testing and purposive testing.

Representative testing is concerned more with the quality of language. One might assume that a student's ability to correctly fill a **cloz**e passage means the student grasps the structure of the language. A *cloze test* (also called *cloze deletion test*) is an exercise, test, or assessment consisting of a portion of text with certain words removed (cloze text), where the participant is asked to replace the missing words. Cloze tests require the ability to understand context and vocabulary in order to identify the correct words or type of words that belong in the deleted passages of a text. Purposive testing focuses on effectiveness of communication by students. Can the students write a letter appropriately or follow instructions on a writing exercise?

Diagnostic test is also known as formative or progress test. These tests let the teacher and students know how well they have learnt particular course elements and are done at the end of course book unit or recent class work. The test content and question types should be familiar to students and a high degree of success is usually expected since they know what is in the test. For example, at the end of a lesson on English prepositions, a language teacher can give test items in form of sentences of about twenty (20) with missing prepositions. Students can be asked to fill in the gaps with the appropriate preposition in each sentence. The results of such exercise will show which areas need revising with the class or individuals.

Achievement test is also called attainment or summative test. It aims to measure what has been learnt over a longer period of time than a diagnostic test. An achievement test does not relate to a particular course book but aims at the syllabus. For example, the syllabus covered for the Senior Secondary Certificate Examination in English language is designed to measure achievement test on a larger scale. The syllabus may specify broad range of lexical and structural content, letter writing, comprehension, summary exercises and Oral English. However, a problem with these tests is that you cannot examine everything in the course of a few hours, so you have to choose samples.

According to Maduekwe (2007), achievement test is expected to reflect success not failure. They should reinforce the learning that has taken place, not go out of their way to expose weakness. They can also help in deciding on changes for future teaching programmes where students do significantly worse in (parts of) the test than might have been expected.

Placement test is used to group (place) new students in the right class in a school. The tests are based on syllabuses and materials students will follow. Here, the subject matter of any reading and listening texts, speaking and writing tasks are usually based on common experience --- something everyone can relate to. An interview can be used to find out a student's spoken accuracy and fluency. The Joint Universities, Polytechnics and Colleges of Education Matriculation Examination and the Post-UME

examinations conducted by autonomous universities, polytechnics and colleges of education are examples of placement tests in Nigeria.

Language aptitude test is used to predict a person's success on exposure to the foreign or second language. According to Carroll and Sapon (1958), language aptitude test does not refer to whether or not an individual can learn a foreign language, but it refers to how well an individual can learn a foreign language in a given amount of time and under given conditions. In other words, the purpose of this test is to determine how quickly or easily a learner learns language in a language training programme.

It should be noted, however, that these five test types overlap. There are elements of proficiency in the other four types, but all have to be valid, reliable and practicable

2. Based on orientation, language testing is divided into:
a. Language competence test
b. Performance language test.

Language competence test, according to Dewi and Nastiti (2012), is a test that involves components of language such as vocabulary, grammar and pronunciation while **performance language test** is a test that involves the basic skills in English which are writing, speaking, listening and reading.

Direct competence tests are tests that focus on measuring the student's knowledge about language components like grammar or vocabulary in which the elicitation uses one of the basic skills (speaking, listening, reading, or writing). For instance, a teacher who wants to know the extent of student's knowledge of grammar can ask them to write a narrative essay to evaluate them. On the other hand, **indirect competence tests** are tests that focus on measuring student's knowledge about language components like grammar or vocabulary in which the elicitation does not use one of the basic skills; speaking, listening, reading or writing. Other ways are used like the multiple choice method where a teacher gives a multiple choice test to students to know their grammar knowledge.

Performance language test can be direct or indirect. **Direct performance test** is a test that focuses on measuring the students' skill in reading, writing, speaking and listening and the elicitation is through direct communication. For example, if a teacher wants to know the students' skill in writing, he can ask them to write a letter or an essay. On the other hand, **indirect performance test** is a test that focuses on measuring the students' skill in reading, writing, speaking and listening and the elicitation does not use any of these skills.

The Concept of Reliability in Language Testing

The characteristics of a good language test are; reliability, validity, practicality and fairness but the focus of this paper is predominantly on 'reliability in language testing'

'Reliability' is one of the most important characteristics of all tests in general, and language tests in particular. In fact, an unreliable test is worth nothing. In order to understand the concept, an example may prove helpful. Suppose a student took a test on grammar comprising one hundred (100) items and got a score of 90. This student further took the same test two days later and got 45. For the third time, the student took the same test items and received a score of 70. Since the student's knowledge of grammar cannot go under drastic changes within this short period, the best explanation would be that there must have been something wrong with the test items administered to the student. The test is, therefore, not reliable because it does not produce consistent scores. Hence, it is not possible to make a sound decision on the basis of such test scores. This is the essence of the concept of reliability, that is, producing consistent scores.

According to Nunally (1982), reliability is concerned with the extent to which measurements are repeatable if all items being studied were included. In other words, reliability can be described as the extent to which a test measures what it purports to measure consistently and accurately. In the same vein, Maduekwe (2007) stated that test reliability refers to the idea that a good language test should give consistent results. A reliable English test, in her opinion, is one which should measure whatever it is supposed to measure consistently under all conditions. For example, if the teacher administers three tests in English language class say for a term, and the students perform in a consistent manner on the tests, then the test items are reliable. Also, Bachman and Palmer (1996) defined reliability as "the consistency of measurement". That is, a language test is reliable to the extent that whatever it measures, it measures it consistently. A measure is considered reliable if a person's score on the same test given twice is similar. On his part, Jacob (1991) stated that "reliability is an essential characteristic of a good test, because if a test doesn't measure consistently (reliably), then one could not count on the scores resulting from a particular administration to be an accurate index of students' achievement". Reliability, therefore, shows the extent to which test scores are free from errors.

Types of Reliability

There are four major methods of determining reliability in language testing. They are namely; test-retest, parallel forms, inter-rater and item reliability.

Test-retest reliability is a measure of reliability obtained by administering the same test twice over a period of time to a group of individuals. The scores obtained in the two tests can then be correlated in order to evaluate the test for stability over time.

Parallel or alternate form reliability is a measure of reliability obtained when a language teacher creates two forms of the same test by varying the items slightly. Reliability is stated as a correlation between scores of Test 1 and Test 2.

Inter-rater reliability is a measure of reliability used to assess the extent to which different judges or raters agree in their assessment decisions. This is because two teachers will not necessarily interpret answers the same way. A teacher's mark for the same composition exercise may vary in accordance with his physical condition, emotional state or any other circumstance that might affect his judgment. It is possible he may give a different mark if he were to assess the same composition after an interval of a month or a term.

Item reliability is a measure of reliability used to evaluate the degree to which different test items that probe the same construct produce similar results. This is because the test items may not be reliable. The items, for instance, may be too easy or too difficult and the items may not discriminate sufficiently between intelligent and not too intelligent students.

Factors which Affect Reliability of Language Tests

The major factor which affects reliability of language test items is the length of the assessment. Sattler (2001) stated that test length is a major factor in reliability of tests. The longer the test is, the more reliable it is. Moreover, in practice, reliability is enhanced by making the test instructions absolutely clear, restricting the scope for variety in the answers and making sure the test items remain constant.

Hughes (2003) gave two reasons why tests are unreliable. The first is the interaction between the person taking the test and the features of the test itself. Human beings are not machines and, therefore, should not be expected to perform in exactly the same way on two different occasions in whatever test they take. As a result, expect some variation in the scores a person gets on a test, depending on when he/she took it, the mood of the person when he/she took the test, how much sleep the person had before the exercise. However, the language teacher who administers the test can ensure that the tests themselves do not increase this variation by having unclear instructions, ambiguous questions, or items that result in guessing on the part of the test takers or students. Unless, there is a deliberate attempt to minimize these features, the scores students obtain on such test cannot be judged as reliable.

Secondly, Hughes (2003) stated that the scoring of a test is another source of unreliability. According to him, scoring can be unreliable in that equivalent test performances are accorded significantly different scores. This is the case in inter- rater reliability where the same composition may be given very different scores by different

markers or even by the same marker on different occasions. Most large testing organizations like the West African Examination Council (WAEC), National Examination Council (NECO) and National Board for Technical Examination Board (NABTEB), take every precaution to make their tests and the scoring of them, as reliable as possible. They are generally highly successful in this respect. They achieve this feat by providing a uniform marking guide/scheme in all the subjects examined including English language and co-ordination meetings are held by the chief examiners (CE) and team leaders (TL), then followed by the assistant examiners (AE) in each marking centre. There, dummy scripts are marked and the apportionment of marks across the questions is discussed as stipulated in the marking guide. This ensures reliability of scores in the subjects taken by the students. Hence, inter-rater reliability is minimized or reduced to the barest minimum. On the other hand, small-scale testing tends to be less reliable. These include classroom tests conducted during a term/semester or end of term/semester or session examinations taken in post-primary schools and tertiary institutions where a uniform marking guide/scheme may not be used by the language teachers teaching different classes.

In addition, reliability can be problematic when a test is a speed test because it is not every student that is able to complete all the items in a speed test. In contrast, a power test should .be used in which every student is able to complete all the test items.

Also, group homogeneity and item difficulty are important factors. This is because the more heterogeneous the group of students who take the test, the more reliable the measure will be and the extent or degree of item difficulty will lead to test scores being reliable or unreliable. Reliability will be low if a test is so easy that every student gets most or all of the items correct or so difficult that every student gets most or all of the items wrong. Hence, when there is little variability among test scores, the reliability will be low.

Aspects of the testing situation can also have an effect on reliability. For example, if the test is administered in a room that is extremely hot, respondents might be distracted and unable to complete the test to the best of their ability. Other things like fatigue, stress, sickness, poor instructions, motivation and environmental distractions can have influence on the reliability of the measure.

In conclusion, the test-retest interval should not be too long. The shorter the interval between two administrations of a test, the less likely that changes will occur and the higher the reliability of test scores obtained.

Conclusion and Recommendations on How to Improve Reliability of Language Tests

The following suggestions will go a long way to make language testing more reliable if the procedures are adopted by English language teachers and test administrators.

Firstly, the language teacher should pay more attention to the careful construction of the test items. Phrase each item clearly so that students know exactly what they are requested to do. Try to write items that discriminate among good and poor students and are of an appropriate difficulty level. The questions should neither be too easy nor too difficult for the students.

Secondly, longer tests tend to reduce the influence of chance factors such as guessing. So, essay questions are preferable to multiple choice questions because the latter requires more time to write than the former. Setting longer tests improves reliability only when the additional items are of good quality and as reliable as the original ones. Adding poor quality items induces error and lowers reliability.

Thirdly, the English language teacher should start planning the test and writing the items well ahead of the time the test is to be given to students. If this is not done, he is likely to write the test items hurriedly at the last minute and this will make the test have low reliability.

In addition, the teacher should construct the test items using clear instructions and directions. Poorly worded or ambiguous items or trick questions are another major threat to reliable measurement.

Also, to improve the reliability of language test scores, the scorers, raters or examiners should score objectively and not subjectively. An objective test is more reliable because the test scores reflect true differences in achievement among students and not the judgment and opinions of the scorer.

Moreso, students should be identified by number, not name. This is because, scorers, most of the time, have expectations of candidates that they know and this affects the way that they score, especially in subjective marking like compositions. The identification of the candidates only by numbers will reduce such bias on the part of the raters or scorers.

Finally, if all these suggestions are taken, language teachers will find it less difficult to identify the problems of students in language learning situation and the reliability of language test scores will be enhanced to nearly 100% (percentage).

References

Bachman, F.L & Palmer, A.S. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Carroll J. & Sapon S. (1958). *The Modern Language Aptitude Test*

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. MIT: M.I.T Press

Dewi, R.K & Nastiti N.S. (2012) *Kinds of Test* http://thejoyoflanguageassessment. wordpress.com/2012/12/19/kindoftest/

Hughes, A. (2003) *Testing for Language Teachers* 2nd edition. Cambridge: Cambridge University Press.

Hornberger, N.H., Shohamy, E. (2008) *Encyclopedia of Language and Education* Volume 7 in Language Testing and Assessment. Berlin: Springer ISBN 0-387-32875-0

Jacobs, L.C. (1991) *Test Reliability http://www.indiana.edu/_best/bweb3/test-reliability*

Maduekwe, A.N. (2007) *Principles and Practice of Teaching English as a* Second *Language.* Lagos: Vitaman Educational Books.

Nunally, J.C. (1982) "*Reliability of Measurement*" *Encyclopedia of Educational Research (*4) pp 15-16

Phelan, C. & Wren, J. (2005) *Exploring Reliability in Academic Assessment* http://www.uni.edu/chfasoa/reliabilityandvalidity.htm

Sattler. J.M. (2001) *Assessment of Children Cognitive Applications* 4th Edition. USA: Publisher Inc.