UNDERSTANDING THE VALIDITY AND RELIABILITY OF ASSESSMENT TESTS

Nguyen Luong Hoang Thanh, M.A.

ABSTRACT

It has long gone without saying that assessment has a crucial role to play in the process of teaching and learning of every subject. Another common knowledge is that designing an assessment procedure which appropriately demonstrates student capacity is not in the least easy. Among numerous characteristics of a well-designed test, validity and reliability should be attentively taken into account. If we are to interpret the score on a given test as an indicator of an individual's ability, that score must be both valid and reliable (Bachman, 1990, 24). These two concepts are not only essential when analyzing and using measures of language abilities but also foremost to be considered when developing and giving tests. In light of this, the research paper will discuss in depth the validity and reliability of assessment tests as well as the intimate relationship between these two qualities.

Keywords: validity, reliability, assessment

INTRODUCTION

First of all, I would like to have an overview about the purpose and the role of assessment. As for the purpose of assessment, Snow, in his More Than a Native Speaker (2006) has claimed that test and other evaluative measures come in a variety of colors; tests alone can be broken down into 4 or more main categories (for further discussion, see Bailey, 1998, 37-39; Harmer, 2001, 321; Hughes, 1989, 9; Madsen, 1983, 8-9). However, the underlying purposes of almost all classroom evaluation fall into two main categories. Diagnostic: one reason to evaluate is to determine how well students are doing in their studies. This information helps students assess how much progress they are making and where they are weak and strong. It also helps you determine how effectively a course facilitates student learning. Motivational: the most obvious motivational effect of evaluation on students is the incentive it gives them to study harder. Talking about the role of assessment, we can take the importance of this for granted. In the journal of Prodromou (1995), it goes without saying that tests and examinations – at the right time, in the right proportions – have a valuable contribution to make in assessing learners' proficiency, progress and achievement. They are dispensable not only for the learners themselves but also for the teachers. Madsen (1983, p 4-

5) shared the same viewpoint. For students: well made tests of English can help students in at least two ways. First of all, such tests can help create positive attitudes toward your class. A second way that English tests can benefit students is by helping them master the language. In short, properly made English tests can help create positive attitudes toward instruction by giving students a sense of accomplishment and a feeling that the teacher's evaluation of them matches what he has taught them. Good English tests also help students learn the language by requiring them to study hard, emphasizing course objectives, and showing them where they need to improve. For teachers, tests help us answer the important question "Have I been effective in my teaching?" In other words, we can use them to diagnose our own efforts as well as those of our students. We might well ask ourselves the following questions: "Are my lessons on the right level? Am I teaching some skills effectively but others less effectively? What areas do we need more work on? What points need reviewing?" Acknowledging the important purposes and roles of assessment, I will continue by introducing the test of my English center. According to Davies (1990), in terms of tests use and test purpose, we can distinguish at least 5 uses: achievement, proficiency, aptitude, diagnosis and pre-achievement. The one our school has been using is achievement test. In this particular kind of test, the concern is with the measuring what has been learnt of what has been taught or what is in the syllable, textbook, materials, …

LITERATURE REVIEW

If we are to interpret the score on a given test as an indicator of an individual's ability, that score must be both reliable and valid. These qualities are thus essential to the interpretation and use of measures of language abilities, and they are the primary qualities to be considered in developing and using test (Bachman, 1990, 24).

First of all, I will discuss validity. Talking about validity, we address the most important question of all in language testing: "Does the test test what it is supposed to test?" This issue should be of central concern to all testers, since if a test is not valid for the purpose for which it was designed, then the scores do not mean what they are believed to mean (Anderson, Clapham and Wall, 1995). Madsen (1983, p178-179) and Harmer (2007, p381) also made a similar claim and some examples. For Madsen, a valid test is one that in fact measures what it claims to be measuring. A listening test with written multiple choice options may lack validity if the printed choices are so difficult to read that the exam actually measures reading comprehension as much as it does listening

comprehension. It is least valid for students who are much better at listening than at reading. Similarly, a reading test will lack validity if success on the exam depends on information not provided in the passage for example, familiarity with British or American culture. And Harmer, in his turn, thought that a test is valid if it tests what it is supposed to test. Thus it is not valid, for example, to test writing ability with an essay question that requires specialist knowledge of history or biology – unless it is known that all students share this knowledge before they do the test. Besides, Bachman (2004, p259-260) has discussed there are several aspects of the conceptualization of validity that are important to understand and keep in mind. Linn and Granlund (2000: 75-6) list the five "cautions" in the use of the term validity. For Thorndike and Hagen (1986), over recent years, the increasing interest in different aspects of validity has led to a confusing array of names and definitions, but most testers, even if they have used different terms, have identified 3 main types of validity: rational, empirical and construct validity. However, according to Anderson et al. (1995), as research into test validity has progressed, it is no longer useful to make the rational/empirical distinction, since both methods of validation may include empirical data. People therefore use the terms internal and external validity, with the distinction being that internal validity relates to studies of the perceived content of the test and its perceived effect, and external validity relates to studies comparing students' test scores with measures of their abilities of learned from outside the test. Internal validity incorporates face validity, content validity and response validity. External validity involves concurrent validity, predictive validity and consequential validity.

While validity is the most important quality of test interpretation or use, reliability is a quality of test scores themselves (Bachman, 1990). It is often defined as consistency of measurement. A reliable test score will be consistent across different characteristics of the testing situation. Thus, reliability can be considered to be a function of the consistency of scores from one set of tests and test tasks to another (Bachman and Palmer, 1996). Other linguists have explored the same idea. These are their arguments: if a guided oral interview were being administered on 2 occasions, reliability would probably be hampered if the teacher on the first occasion was warm and supportive and the teacher on the second occasion abrupt and unfriendly (Madsen, 1983, p178-179). If the same group of students took the same test twice within two days without reflecting on the first test before they sat it again – they should get the same results on each occasion. If they took another similar test, the results should be consistent. If two groups who were demonstrably

alike took the test, the marking range would be the same (Harmer, 2007, p381). As for Brown (2004), we can divide reliability into four types: student related reliability, rater reliability (inter rater reliability and intra rater reliability), test administration reliability, and test reliability. According to Hughes (1989), there are two components of test reliability: the performance of candidates from occasion to occasion and the reliability of the scoring. For the first one, Hughes has indicated some methods such as: test retest methods, alternate form method, and split half method. For the latter one, the most famous method is inter rater or inter observer. If the 2 raters/observers gave widely different ratings to the same test, the result of the test would be unreliable.

What is the relationship between the two concepts? Most linguists agree that there is a close relationship and both of them are essential to the use of tests. However, Alderson et al. (1995) have claimed that in principle, the relationship is a simple one but in practice rather complex and not well understood. In principle, a test cannot be valid unless it is reliable. On the other hand, it is quite possible for a test to be reliable but invalid. Take MCQ as an example, they can be made highly reliable, especially if they contain enough items, yet some testers would argue that performance on a multiple-choice test is not a highly valid measure of one's ability to use language in real life. The problem for most language testers is that in order to maximize reliability, it is often necessary to reduce validity so they should learn how to make a trade off. Which one we should scarify depends on what exactly we are trying to measure by setting the task. As Hughes (1989) has stated, there will always be some tension between reliability and validity. The tester has to balance gains in one against losses in the other.

RECOMMENDATIONS

Hughes (2003, p33-34/44-51) has explored some methods to make the test more valid and reliable. Building upon these findings, the researcher proposes strategies to improve the validity and reliability of tests at our school. To enhance the face validity, students' understanding of the test's importance should be emphasized. By recognizing the test as a measure of progress, students will be motivated to perform their best. Besides, encouragement in the form of scholarships or gifts can further incentivize their efforts. We should also provide a test with a variety of testing items so that the students do not have to do the same thing from one task to another and they will feel less bored when sitting the exam. This also helps to evaluate different skills of students such as

analysis skill or synthesis skill. A variety of testing items also permits more objective scoring, for example we can use True/False, Matching, Multiple choice questions … For the content of the test, we should revise it so that the test can cover all the important points of the course, be free of spelling mistakes and exclude items which are outside the course. During the exam, the students should be instructed clearly and carefully what they have to do. If the test takers are at the beginning level, instructions should be bilingual. Besides, they need to have a chance to get used to the format of the test as well as the testing techniques some time before the final exam so that they shouldn't be confused or nervous. As for the marking of the test, the teachers should be given more time for marking all the test paper. Besides, reliability can be greatly enhanced by having more than one scorer. Another thing is that some teachers mark the test strictly while others do the task leniently, so we must trainers scorers so that they can mark the same in every situation (ex: wrong tense, wrong word, lack of final sound, …).

REFERENCES

Alderson, J. C., Clapham, C & Wall, D. 1995. *Language Test Construction and Evaluation.* Cambridge: CUP.

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing.* Oxford: OUP.

Bachman, L. F. 2004. *Statistical Analyses for Language Assessment.* Cambridge: CUP.

Bachman, L. F. & Palmer, A. S. 1996. *Language Testing in Practice; Designing and Developing Useful Language Tests*. Oxford: OUP.

Brown, H. D. 2004. *Language Assessment - Principles and Classroom Practices*. White Plains, NY: Longman.

Davies, A. 1980. *Principles of Language Testing*. Oxford: Basil Blackwell.

Harmer, J. 2007. *The Practice of English Language Teaching.* 4th edition. Pearson Longman ELT.

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge: CUP.

Hughes, A. 2003. *Testing for Language Teachers*. 2nd edition. Cambridge: CUP.

Madsen, H. S. 1983. *Techniques in Testing*. Oxford University Press.

Prodromou, L. 1995. *The Backwash Effect: From Testing to Teaching*. ELT JOURNAL 49/1 January 1995.

Snow, D. 2006. *More Than a Native Speaker: An introduction to teaching English abroad*. 2nd edition. Alexandria, VA: TESOL.