ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 2, No. 1, pp. 267-269, January 2011

doi:10.4304/jltr.2.1.267-269

Validity Considerations in Designing an Oral Test

Wei Li
School of Foreign Languages, Henan University of Technology, Lianhua Street, Zhengzhou City, China
Email: gracie.lee@163.com

***Abstract*—A language test is said to be valid if it measures accurately what it is intended to measure. Qualities of tests involve validity considerations such as validity, reliability, authenticity, intractiveness, impact and practicality, which should also be taken into account in designing an oral test.**

***Index Terms*—language test, validity considerations, oral test**

I. Introduction

Whether for educational uses or for researches information about people's language ability is nowadays often very useful and sometimes necessary. It is difficult to imagine, for example, British and American universities accepting students from overseas without some knowledge of their proficiency in English. They certainly need dependable measures of language ability. The same is true for organizations hiring interpreters or translators. Then it comes to the requirement of dependable measures -- the use of language tests. Language tests refer to instruments used to measure language ability or aptitude. A defining feature of language tests is that they consist of "specified tasks through which language abilities are elicited" (Daves, 2002, p. 107). Hughes (2000) believed that "a test is said to be valid if it measures accurately what it is intended to measure"(p. 22). But too often language tests have a harmful effect on teaching and learning; and they fail to measure accurately what they are intended to measure. The effect of language tests on teaching and learning is known as backwash. Backwash can be positive or negative. For example, the use of an oral interview in a final examination may encourage teachers to practice conversational language use with their students. And if a test is regarded as important, then preparation for it can come to dominate teaching and learning activities. Good tests can be supportive of teaching whereas tests with poor quality which lack test techniques and reliability would have harmful effect on teaching and learning. Therefore tests with qualities are necessary.

II. Validity Considerations in Designing an Oral Test

Qualities contributing to designing a valid test involve validity considerations such as validity, reliability, authenticity, interaction, impact and practicality. The same is true with an oral test which aims at testing how efficiently language speakers could interact in that language. These validity considerations should be taken into account in designing an oral test to make it a valid one.

A. Validity

Validity in general refers to the appropriateness of a giving test or any of its component part as a measure of what it is supposed to measure. It is the quality which most affects the value of a test, prior to, though dependent on, reliability. However, validity is related to the content and construct of a test while reliability is related to the score.

The concept of validity reveals a number of aspects as mentioned above, each of which deserves attention. The most commonly referred to types of validity are face validity, content validity, concurrent validity, predictive validity and construct validity. A test is said to have face validity if it looks as if it is supposed to measure.

To achieve face validity, an oral test may use direct method such as picture tasks, dialogues, group discussion, role play, interpreting, imitation or pair work with the attempt to duplicate as closely as possible the setting and operation of the language use situations; meanwhile, direct method makes oral tests authentic for it is reciprocal in nature and there's more interaction between the task and the test taker.

An oral test is said to have content validity only if it includes a proper sample of the relevant structures, whether dialogue, discussion, role play or pair work. While designing an oral test, testers should above all make clear the purpose of it, to set tasks that form a representative sample of the population of oral tasks that candidates are able to perform, whether for educational purpose such as achievement, oral proficiency, etc, or for occupational purpose such as job competency. The tasks should elicit behavior which truly represents the candidates' ability and which can be scored validly and reliably. And test takers' background knowledge, levels of language should also be considered.

Concurrent validity refers to the comparison of the test scores with some other measures for the same candidates taken at roughly the same time as the test intends to. Since it is impractical to reflect test takers' oral proficiency within a refined time, testers may choose at random a sample of all the students taking the oral test and rank their performance, then compare the two rankings–students' test results and their performance. The more similar the two groups of marks are, the higher concurrent validity the oral test has.

Predictive validity measures how well a test predicts performance on an external criterion. An oral test is said to have higher predictive validity if performance on the test correlates highly with performance (e.g. as measured by grades) on a subsequent oral course which is taught through the language under test (Daves, 2002).

The construct validity of a language test is an indication of how representative it is of an underlying theory of language learning. A language test is said to have construct validity if it can be demonstrated that it measures just the ability which it is supposed to measure. "Construct validation involves an investigation of the qualities that a test measures, thus providing a basis for the rationale of a test" (Daves, 2002, p. 33). According to Hughes (2000), "the word 'construct' refers to any underlying ability (or trait) which is hypothesized in a theory of language ability" (p. 33). To establish construct validity of an oral test, one may hypothesise that speaking ability (or construct) involves a number of traits (or subconstructs) such as grammar, accent, fluency, and comprehension. An oral test with those traits constructed may be administered to a group of students as a pilot test. Then testers score them reliably and find out the internal correlations of subconstructs, the relationship between each subconstruct, and the relationship between subconstruct and construct. If coefficients between various traits of speaking ability are low but those between the total construct and each subconstruct are high, the oral test is said to have a high construct validity.

B. Reliability

Reliability is also an absolutely essential quality of tests which means consistency in scores regardless of when and how many times a particular test is taken. The more similar the scores would have been, the more reliable the test is said to be (Hughes, 2000). There are two components of reliability: the performance of candidates from occasion to occasion, and the reliability of the scoring. In the same way, to make an oral test reliable, testers should try out to achieve consistent performances from candidates and to achieve scoring reliability.

Then how to make candidates perform consistently in an oral test? There are some ways of achieving consistent performances from candidates. Firstly, use more items in an oral test, for the more items a test has, the more reliable that test will be. "It has been demonstrated empirically that the addition of further items will make a test more reliable" (Hughes, 2000, p. 36). However, one thing to bear in mind is that the additional items should be independent of each other and of existing items. And each additional item should as far as possible represent a fresh start for the candidate. In an interview used to test oral ability, the candidate should be given as many 'fresh starts' as possible. By doing so additional information on the candidates may be gained which will make the results of the oral test more reliable. Secondly, provide clear and explicit instructions so that candidates can avoid introducing confusion. Thirdly, candidates should be familiar with the format and testing techniques. Thus efforts must be made to ensure that all candidates have the opportunity to learn just what will be required of them. Fourthly, uniform and non-distracting conditions of administration should be provided.

Score reliability is the other essential component to test reliability. Score reliability consideration of an oral test involves scoring and criterial levels. Criterial levels are specified to obtain valid and reliable scoring. For example(Hughes, 2000), in the Royal Society of Arts (RSA) test of the Communicative Test of English as a Foreign Language, criterial levels such as accuracy, appropriacy, range, flexibility, and size are used to evaluate test takers' oral ability. Moreover, the operationalisation of criterial levels has to be carried out in conjunction with samples of candidates' performance. To make an oral test have a higher score reliability, global scoring (or holistic scoring) and discrete scoring (or analytic scoring) method may be adopted. Comparably speaking, discrete scoring is more detailed and more feasible while holistic scoring tends to be general. It is better to use one method as a check on the other. In the end, administration should standardize oral procedures and raters of an oral test should be trained regularly to make it more reliable.

Actually, validity and reliability are quite interrelated with each other but they have different focuses. Validity focuses on test content such as test purpose or use while reliability focuses on the result or response or score. Validity is rather more comprehensive while reliability is much narrower. Reliability is a necessary condition for validity but it alone is not sufficient. Hence a balance should be made between validity and reliability in designing an oral test.

C. Other Validity Considerations

Besides validity and reliability, practicality is often quoted as the third consideration in test design. Its inclusion as a major concern stems from the recognition that however valid and reliable a test may be, if it is not practical to administer it in a specific context then it will not be taken up in that context (Daves, 2002). Practicality means the extent to which the demands of the particular test specifications can be met within the limits if existing resources such as human support (raters, examiners, etc.), technical support (lab facilities, recording facilities, etc.) and logistics support (computers, rooms available, etc.). These practicality considerations should also be taken into account in designing an oral test. If personnel are in short, labs may be used to solve the problem. And it certainly makes sense to make the best use of existing logistics in designing a valid and reliable oral test.

Still other aspects are important for a valid oral test. Authenticity refers to the extent or degree of correspondence between characteristics (form and skill) of TLU (target language use) tasks and those of test tasks. As mentioned above, direct testing method may be used in an oral test such as picture tasks, dialogues, group discussion, role play, interpreting, imitation or pair work to make it authentic and interactive.

Interactiveness means the extent and types of involvement of the test takers' individual characteristics in

accomplishing a test task. If a face-to-face interview procedure is adopted in an oral test, it would be highly interactive for it involves a lot of interaction (comprehension as well as production) between testers and testees and would have a close resemblance to real language behavior.

Impact refers to the extent or degree to which a particular test influences society and educational systems and the individuals within these systems. There can be high-stake tests and low-stake tests. High-stake tests' scores will have great influence on test takers, schools, institutions and even the society such as CET. A test may have a significant impact on the career or life chances of individual test takers, particularly if it has a gatekeeping function. Other stakeholders (e.g. teachers, employers, course admissions officers) may also be affected by the introduction, administration or results of a test. A test can affect society on a larger scale when used to make decisions about, for example, immigration (e.g. oral interview in IELTS), certification for professional practice or the amount and kind of instruction to be given to school children (Daves, 2002). Test makers certainly hope an oral test to have a positive effect or beneficial backwash on pedagogue and speakers' ability as well as the society.

III. Summary

The accurate measurement of oral ability is not easy. It takes considerable time and effort to obtain valid and reliable results. Nevertheless, where backwash is an important consideration, the investment of such time and effort may be considered necessary. The appropriateness of content, descriptions of criterial levels, and elicitation techniques used in oral testing should depend upon the needs of individual institutions or organizations.

Validity considerations mentioned above are indispensible in designing an oral test. Specifically, oral test makers should set tasks that form a representative sample of the population of oral tasks that they expect candidates to be able to perform. The tasks should elicit behavior which truly represents the candidates' speaking ability and which can be scored validly and reliably. In that way could testers design a valid oral test with the expectation of a beneficial impact on teaching and learning and on the society.

References

[1] Alderson, C. et al. (2000). Language Test Construction and Evaluation. Beijing: Foreign Language Teaching and Researching Press.
[2] Bachman, Lyle F. & Adrian S. Palmer. (1996). Language Testing in Practice. Shanghai: Shanghai Foreign Language Education Press.
[3] Brown, J.D. (2001). Understanding Research in Second Language Learning. Beijing: Foreign Language Teaching and Researching Press.
[4] Daves, A. et al. (2002). Dictionary of language Testing. Beijing: Foreign Language Teaching and Researching Press.
[5] Heaton, J.B. (2000). Writing English Language Tests New Edition. Beijing: Foreign Language Teaching and Researching Press.
[6] Henning, Grant. (2000). A Guide to Language Testing. Beijing: Foreign Language Teaching and Researching Press.
[7] Hughes, Arthur. (2000). Testing for Language Teachers. Beijing: Foreign Language Teaching and Researching Press.
[8] Shen Zou. (2005). Language Test. Shanghai: Shanghai Foreign Language Education Press.

Wei Li was born in Zhumadian, China in 1971. She received her M.A. degree in English language and literature from Central China Normal University in Wuhan City, Hubei Province, China in 2003.

She is currently an associate professor in the School of Foreign Languages, Henan University of Technology, Zhengzhou City, Henan Province, China. Her research interests include pragmatics and literature.

Ms. Li has published more than ten academic papers on journals in Canada and China such as Different Communication Rules between the English and Chinese Greetings in *Asian Culture and History* and Different Interpersonal Relationships Underlying the English and Chinese Greetings in *Asian Social Science*.