Research Journal of English Language and Literature (RJELAL)
A Peer Reviewed (Refereed) International Journal
Impact Factor 6.8992 (ICI) http://www.rjelal.com;
Email:editorrjelal@gmail.com ISSN:2395-2636 (P); 2321-3108(O)

Vol.8.Issue 2. 2020 (April-June)

RESEARCH ARTICLE

INTERNATIONAL STANDARD SERIAL NUMBER INDIA
2395-2636 (Print):2321-3108 (online)

MAJOR LANGUAGE TEST QUALITIES AND WAYS OF ENHANCING THEM

NGUYEN NGOC THUY[1], NGUYEN LUONG TUAN DUNG[2]

[1,2]Lecturer, Ho Chi Minh city University of Natural Resources and Environment, 236B Le Van Sy, Ward 1, Tan Binh District, Ho Chi Minh city, Vietnam

[1]Email: toeic05dh.av11@gmail.com'[2]Email: dung.nlt@hcmunre.edu.vn

Article Received:02/03/2020
Article Accepted: 24/03/2020
Published online: 07/04/2020
DOI: 10.33329/rjelal.8.2.1

Abstract

According to Brown (2004), constructing a good test is a complex task that involves both science and art. Since a test draws on a limited sample of observable behaviors, it sometimes fails to reflect the test-taker's true ability. Other factors may also cause the inaccuracy of tests scores. Developing "good" tests is, therefore, very crucial not only for teachers but also the educational system in general. This article begins with a general introduction of language testing and language test qualities. It also discusses major qualities that are necessary for any good language tests, the reliability and validity of a test, and explains why they are important simultaneously. Moreover, suggestions on how to improve these qualities in English tests are provided, which includes validity, reliability, washback and authenticity. To make the matter clearer, examples are also given to illustrate the points. The article hopes to provide English teachers, as well as language test designers with concise knowledge of language tests, and ways of enhancing their tests, in order to improve their tests and their work.

Keywords: test quality, good test, validity, reliability, washback.

INTRODUCTION

Qualities of language test

As "language testing and assessment at any level is a highly complex undertaking that must be based on theory as well as practice" (Christine Coombe & Nancy Hubley, cited by Le Hoang Dung, 2016), many authors provide different principles which can be used to evaluate the test designing result. According to Bachman & Palmer (1996), test usefulness is "the essential basis for quality control throughout the entire test development process" and can be constructed by six elements: *reliability, (construct) validity, authenticity, interactiveness, impact*, and *practicality*. To share something in common to Bachman & Palmer (1996), Brown (2001) identified just "five cardinal criteria" for "testing a test": *practicality, reliability, validity, authenticity*, and *washback,* but not the *interactiveness* as a criterion to recheck a test.

To approach the problem from a different angle, Hughes (2003) lists the requirements for "every test" by combining these criteria into three groups. Firstly, it must consistently provide accurate measures of precisely the abilities in which we are interested (*validity* and *reliability*). Secondly, it must have a beneficial effect on teaching in those cases where the test are likely to influence teaching (*washback*). Finally, it must be economical in terms of time and money (*practicality*) (Hughes, 2003 as cited by Sarosdy et al, 2006). In a brief view, Le Hoang Dung (2016) affirmed: "high quality classroom assessments (of which tests are one type) provide *reliable*, *valid* and *useful* measures of student

performance." However, Bachman and Palmer (1996) also emphasized the need for test-makers to avoid considering any of the above test qualities independently of the others. Trying to maximize all of these test qualities may lead to conflict, so the test-makers should consider attaining an appropriate balance of these qualities for each specific test. In other words, there is no perfect language test, much or less one test is only perfectly suitable for each testing situation.

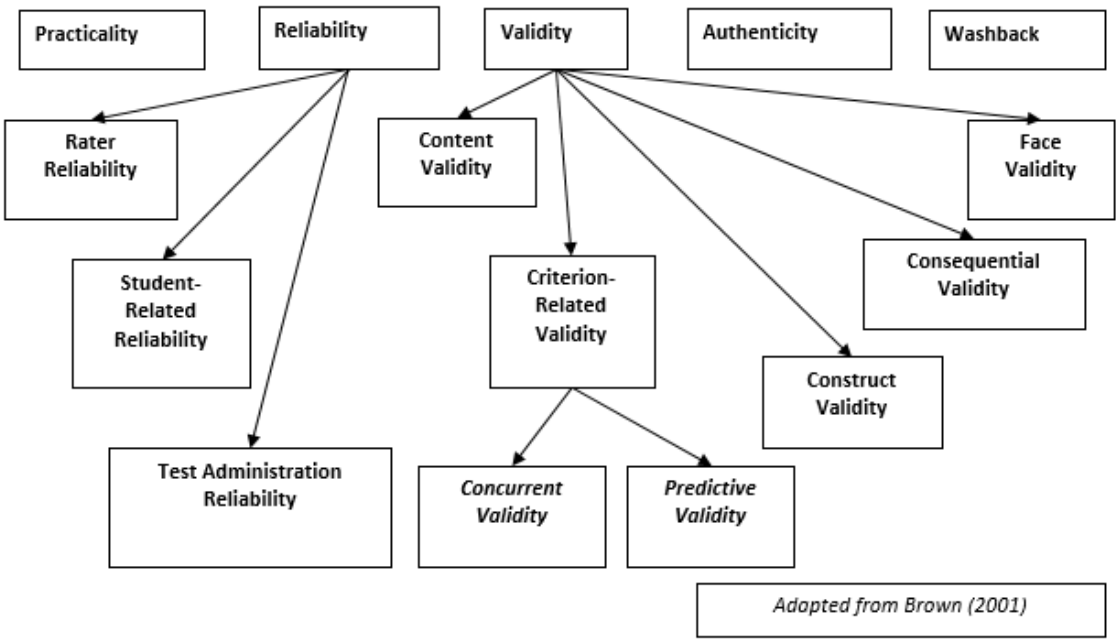

Figure 1. Elements to check the test usefulness

In that spirit, in the first part, the journal mainly focuses on *validity* and *reliability*, the two most important features of a test; since they are mutual concerns of many test-designers as well as researchers. However, in the second part, the research would add *washback* and *authenticity* as criteria should be improved in order to enhance test qualities.

DISCUSSION ON MAJOR QUALITIES OF A LANGUAGE TEST

Validity

Validity is a complex criterion in the field of testing. Validity is the most important issue in testing a test due to three main reasons. First, validity refers to what characteristic the test measures and how well the test measures that characteristic. Besides, validity gives meaning to the test scores. Validity evidence indicates that there is linkage between test performance and language competent of test-takers. It can tell you what you may conclude or predict about someone from his or her score on the test. If a test has been demonstrated to be a valid predictor of performance, you can conclude that people scoring high on the test are more likely to perform well in a language than people who score low on the test, all else being equal. Last but not least, validity describes the degree to which you can make specific conclusions or predictions about people based on their test scores. In other words, it indicates the usefulness of the test.

Another characteristics of validity should be noted by the test-makers is a test's validity is established in reference to a specific purpose; the test may not be valid for different purposes. For example, the TOEIC test is used to make valid predictions about someone's ability to communicate in working environment, yet it may not be valid for predicting his or her working or leadership skills. Test's validity is established in reference to specific groups. A test to predict the performance of students in English for Computing may not make valid or meaningful predictions about the

performance of Business English learners, for instance.

In short, validity of a test covers the appropriateness of that test basing on the purpose of producing. A valid test must meet the requirement of fulfilling the aims of checking to the right group.

Reliability

Bachman and Palmer (1996:19) define reliability as "consistency of measurement". To make it easier to grasp, Brown stated that If the same group of subjects takes the same test on two different occasions, results should be similar, both in individual scores, and in the rank order within the group (H.D. Brown, 2001: 386; Bachman & Palmer, 1996: 20). If the same written answer in a test is scored by two different markers, the two different scores should be similar (Bachman, 1990: 24). If two forms of the same test are created which are intended to be used interchangeably, an individual should obtain very similar scores on both versions (Bachman & Palmer, 1996:20).

It is important to be concerned with a test's reliability for two main reasons. First, reliability provides a measure of the extent to which an examinee's score reflects random measurement error. Measurement errors are caused by one of three factors: (1) examinee-specific factors such as motivation, concentration, fatigue, boredom, momentary lapses of memory, carelessness in marking answers, and luck in guessing, (2) test–specific factors such as the specific set of questions selected for a test, ambiguous or tricky items, and poor directions, and (3) scoring-specific factors such as non-uniform scoring guidelines, carelessness, and counting or computational errors. These errors are random in that their effect on a student's test score is unpredictable– sometimes they help students answer items correctly while other times they cause students to answer incorrectly. In an unreliable test, students' scores consist largely of measurement error. An unreliable test offers no advantage over randomly assigning test scores to students. Therefore, it is desirable to use tests with good measures of reliability, so as to ensure that the test scores reflect more than just random error. The second reason to be concerned with reliability is that it is a precursor to test validity. Reliability is in fact a prerequisite to validity in performance assessment in the sense that the test must provide consistent, replicable information about candidates' language performance (Clark, 1975). That is, no test can achieve its intended purpose if the test results are unreliable. If test scores cannot be assigned consistently, it is impossible to conclude that the scores accurately measure the domain of interest. Validity refers to the extent to which the inferences made from a test (i.e., that the student knows the material of interest or not) is justified and accurate. Ultimately, validity is the psychometric property about which we are most concerned. However, formally assessing the validity of a specific use of a test can be a laborious and time-consuming process. Therefore, reliability analysis is often viewed as a first-step in the test validation process. If the test is unreliable, one need not spend the time investigating whether it is valid – it will not be. If the test has adequate reliability, however, then a validation study would be worthwhile.

SUGGESTIONS ON WAYS OF ENHANCING TEST QUALITIES

Validity

First of all, since a test's validity is established in reference to specific groups, the test developers have the responsibility of describing the reference groups used to develop the test. The manual or test specifications should describe the groups for whom the test is valid, and the interpretation of scores for individuals belonging to each of these groups.

Besides, the test-makers must determine if the test can be used appropriately with the particular type of people they want to test. A test's validity is established in reference to specific groups, called the "reference groups" and the group of people for whom the test is used is called "target population" or "target group". The target group and the reference group do not have to match on all factors; they must be sufficiently similar so that the test will yield meaningful scores for other groups. For example, a writing ability test developed for use with college seniors may be appropriate for measuring the writing ability of officers in an English-

for-adult class, even though these groups do not have identical characteristics. In determining the appropriateness of a test for target groups, the test-makers should consider factors such as occupation, reading level, cultural differences, and language barriers.

In addition, in order to make tests more valid, we should well apply several activities to enhancing construct, content validity as well as score validity, which are shown below:

Construct Validity: Before writing a test, test-makers should clearly define the construct or underlying traits that they want to test. A "vocabulary" test could mean a test of word meanings to one teacher, but it may include the knowledge of word parts and the part of speech to another. However, the name "part of speech" may be a part of a "grammar" test for others. A shared view based on a theory of language learning can help increase the construct validity of the test.

If direct testing is practicable, test-makers should try to use it as much as possible. A test is direct when it requires the test-takers to perform exactly the skill which we wish to measure (Hughes, 2003). The so-called "speaking" test that requires the test-takers to choose or write the correct answers instead of saying them has lower construct validity than the one that requires actual speaking. However, scorer reliability may be an unavoidable problem in many direct tests. Therefore, it is essential to devise a valid test first and then try to establish ways of increasing its reliability (Heaton, 1988).

Content Validity: To strengthen content validity, test-makers should always follow the explicit test specifications, which are published by the authorities at university. All course objectives and the contents that need to be tested are listed in the specification. After finishing a test, test-makers should compare the actual test against the list to see how representative the test items are. For example, if an essay writing course aims to teach how to write introduction, body, and conclusion, yet the test includes only the introduction and the conclusion, the content validity of this test is affected.

Validity in Scoring: Both the test-maker and test-markers should make sure that the scoring of answers relates directly to what is being tested. If an interview is given to see how well the students can speak, their ruffled hair styles or clumsy behaviors, although some teachers may regard them as the signs of students' poor preparation, should not be judged because they have nothing to do with English.

Reliability

According to Sarosdy et al. (2006), there are two opponents of test reliability: the reliability of scores on the performance of candidates from occasion to occasion, which can be ensure by the construction and the administration; and the reliability of scoring. (Sarosdy et al; 2006: 135). Therefore, in order to enhancing test reliability, we should look at test construction, administration, as well as scoring reliability.

Test administration: Generally, test reliability is more based on test administrations than test designing. Thus, the following activities should be applied to create a better test.

First, test administration should be strict to prevent students from cheating. The examiners should not give the test-takers too much freedom, they may take advance of it to cheat or copy from each other result. The fact that students do the test individually will help the teachers to mark them more reliably. Second, test-makers must ensure that tests are well laid out and perfectly legible. Poor photocopying can dramatically reduce the quality of a well-written test. A colorful graph or picture can simply be different shades of grey that cannot be understood by the students. Third, students should be provided with uniform and non-distracting conditions of administration. When students with the same listening ability take a listening test in a quiet room, they normally do better than taking the same test is a noisy place. Non-distracting conditions such as well-lit, cool and quiet are important for every kind of tests. Whenever there are more than one test rooms, the room conditions must be the same.

Test construction

a. Test length: The test should be appropriate in length in order to take enough samples of behavior. In general, longer tests produce higher reliabilities. The test must have enough items for a teacher to tell whether the students know the materials. To increase the test reliability, however, additional items should represent a fresh start or be independent of each other and of existing items. In other word, the ability to answer the next question must not depend on the ability to answer the previous question. Otherwise, there is practically no additional question for the student, which means the teachers do not get an additional sample of the students' behavior, so the reliability is not increased. On the other hand, a test should not be made so long that the students become so bored or tired that the behavior they exhibit becomes unrepresentative of their ability.

b. Item quality: Teachers should exclude items which do not discriminate well between weaker and strong students. Items on which strong students and weak students perform with similar degrees of success contribute little to the reliability of a test. Item quality has a large impact on reliability in that poor items tend to reduce reliability while good items tend to increase reliability. How does one know if an item is of low or high quality? The answer lies primarily in the item's discrimination. Items that discriminate between students with different degrees of mastery based on the course content are desirable and will improve reliability. An item is considered to be discriminating if the "better" students tend to answer the item correctly while the "poorer" students tend to respond incorrectly. For examples, multiple-choice tests allow the calculation of the discrimination index (D), which ranges from 1 to 0 to -1. The higher the D, the better the item discriminates. Items with minus D should be excluded from the test.

Besides, the test tasks should not allow the test-takers too much freedom. For example, it is more difficult to compare essays on different topics than those on the same topic with specific conditions such as audience, purpose and length. Requiring every test-taker to do the same well-defined tasks will help the teachers to mark them more reliably.

Moreover, teachers should write unambiguous items. An item that can be interpreted in different ways on different occasions means that the item is not contributing fully to the test reliability. Poor English may also cause ambiguity and is a bad model for the students. Having other teachers and native speakers scrutinize the draft will reduce this problem.

The test-makers have to make sure there is only one correct answer for a multiple-choice test. A multiple-choice item that can be answered in different ways on different occasions makes a test less reliable. The key to a great multiple-choice question, however, is a set of terrific distracters. They must be attractive but have less merit than the correct answer (Salkind, 2013).

The instructors should make candidates familiar with format and testing techniques. When the formats or testing techniques are new, explain them in class before the test. Some students may not know about penalty for wrong guesses in certain multiple-choice tests. They would do better if they knew it.

The reliability of scoring

In order to ensure the reliability of scoring, test-makers should use items that permit scoring which is as objective as possible. Test techniques such as multiple-choice, matching, and true-false do not have a problem with scorer reliability as they do not require judgment from the scorers, making them popular because they are easy and fast to mark. However, they not only encourage guessing, but also do not require the test-takers to produce the language. An alternative is the open-ended item which has a unique, possibly one-word, correct response which the test-takers produce themselves. Having the students write the correct form of the given verbs will certainly better demonstrate their ability to conjugate those verbs than having them do a multiple-choice test.

Together with a test, test-makers should provide a detailed scoring key. A test cannot be a good test unless its correct and complete key is

provided. Specify all acceptable answers and assign points for partially correct responses for short-answer items. Compile a banding system for a particular group of students for a writing or speaking task.

A test should be employed multiple, independent scoring. Subjective tests should be scored by at least two independent trained scorers.

To conclude, teachers should do everything possible to make the test reliable so that it can be valid. If a test is not reliable, it cannot be valid. Also, it is not for sure that a reliable test may not be valid. For example, a multiple-choice "writing" test may be reliable, but it cannot be a valid test on composition writing.

Washback

In order to achieve beneficial washback and avoid harmful washback, there are several steps that test takers should follow.

Teachers should firstly remember to test the abilities whose development they want to encourage to avoid the tendency of checking the easiest points, not the important points. Students will prepare for the final exam differently if they know they will be interviewed instead of taking a multiple-choice test. Preparation for the interview may help them improve their speaking abilities.

Another way to avoid harmful washback is not to overuse multiple-choice items. Although this test format has many advantages and seems to have no major disadvantages in certain areas such as vocabulary (Nation, 2001) or reading comprehension, objective tests can never test the ability to communicate in the target language (Heaton, 1988). If the teachers always give multiple-choice English tests, which only require the students to recognize the correct answers rather than producing the language, the they should not be surprised at all that the students cannot communicate in English.

Authenticity

To make tests more authentic, before teaching and designing a test, test-makers should always think of what language use their students will be likely to encounter. For instance, it would be more common for students in Business class to make conversation in bank or office context rather than in street or family context. Choosing the materials (reading passages, scripts for a listening test, etc.) from the real-world sources will make the test more authentic because the language can be more natural and contextualized than the language found in many stems written for tests by non-native English teachers. Luckily, with the development of mass media means, nowadays it no longer causes any hindrance for test takers to search for these kinds of materials.

CONCLUSION

In conclusion, it is teachers' responsibility to create a valid and reliable test from which the learners' competence can be assessed most accurately. At the same time, tests should not only be as similar to the real-world tasks as possible but also gives beneficial backwash. All the necessary test qualities should be balanced and maximized to ensure good, useful and effective tests.

REFERENCES

1. Alderson, C., Clapham, C. & Wall, D. (2005). *Language Test Construction and Evaluation. 9th edition*. Cambridge: Cambridge University Press.
2. Bachman, L. F. & Palmer, A. S. (1996). *Language testing in Practice: design and developing useful language tests*. Oxford: Oxford University Press.
3. Brown, H. Douglas (2005). *Language assessment: principle and classroom practice*. Essex: Longman.
4. Carrol, John B. (1968). *Language Testing Symposium: A Psycholinguistic Approach*. Oxford: Oxford University Press.
5. Clark, J. (1975). Theoretical and technical considerations in oral proficiency testing. In S. Jones & B. Spolsky (Eds.), *Language testing proficiency* (pp. 10-24). Arlington, VA: Center for Applied Linguistics.

6. Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment – an advance resource book*. New York: Routledge.

7. Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

8. Gampper, C. (2013). Thammasat Review, Special Issue, 2013. *Improving English Test Qualities.* (pp. 73-83)

9. Heaton, J. (1988). *Writing English language test - new edition*. Essex: Longman.

10. Henning, G. (1987) *A Guide to Language Testing: Development, Evaluation, Research*. Cambridge, MA: Newbury House.

11. Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

12. Hughes, A. (2003). *Testing for Language Teachers – Second Edition*. Cambridge: Cambridge University Press.

13. Kluitmann, S. (2008). *Testing English as a Foreign Language: Two EFL-Tests used in Germany*. MA Thesis, University of Albert-Ludwig.

14. Lawson, A. (2008). *Testing the TOEIC: Practicality, Reliability and Validity in the Test of English for International Communication*. Birmingham: The University of Birmingham, Centre for English Language Studies.

15. Le Hoang Dung (2016). *Lessons of Language Testing*. Ho Chi Minh City: MTESOL program at Ho Chi Minh Open University.

16. Nation, I. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

17. Owen, C. et al (1997). *Testing*. Birmingham: The University of Birmingham, Centre for English Language Studies.

18. Salkind, N.J. (2013) *Test & measurement for People who (think they) hate tests & measurement – second edition*. Thousand Oaks, CA: Sage Publications.

19. Sarosdy, J., Bencze, T. F., Poor, Z., and Vadnay, M. (2006). *Applied Linguistics I: for BA students in English*. Budapest: Bolcesesz Konzorcium.

20. Weir, C. (1990). *Communicative Language Testing.* New York: Prentice Hall.

21. Wells, C. & Wollack, J. (2003). *An Instructor's Guide to Understanding Test Reliability*. Madison: University of Wisconsin, Testing & Evaluation Services.