# REVIEWING TESTS ON ENGLISH SPEAKING 6 AT FACULTY OF ENGLISH, HANOI OPEN UNIVERSITY

***Nguyen Thi Mai Huong****

Date received the article: 02/7/2021
Date received the review results: 04/01/2022
Date published the article: 28/01/2022

***Abstract:*** *This paper reviews the tests on English Speaking 6 at Faculty of English, Hanoi Open University with reference to theoretical and practice-based perspectives in order to evaluate the two important test qualities: validity and reliability and seek ways to improve them. For this aim, the researcher collected the actual speaking 6 tests that are currently in use to assess speaking 6 performance at Faculty of English. In addition, the writer conducted in-depth interviews with the experienced teachers who are directly involved in the speaking assessment. Overall, the study found that the test's validity and reliability were seen to accurately measure the abilities defined in the Speaking 6 construct; however, the reliability of the marking process needs to be addressed.*

***Keywords****: speaking test, validity, reliability, assessment, rating*

## I. Introduction

Testing is an important and integral part of the English learning and teaching process. Testing oral production has become one of the most important issues in language testing since the role of speaking ability has become more central in language teaching with the advent of communicative language teaching (Nakamura [18]). However, assessing speaking is challenging (Luoma [14]). Validity and reliability, as fundamental concerns and essential measurement qualities of the speaking test (Bachman [2]; Bachman & Palmer [3]), have aroused widespread attention.

At Faculty of English, Hanoi Open University, the speaking 6 test is one of the four subtests of the midterm and end-of-term 6 examination of English assessment. The examination uses the face-to-face interview format. This paper aims to ascertain the degree of the reliability and validity of the speaking tests. By analyzing the results of the research, teachers will become more aware of the factors affecting validity and reliability of oral assessments, and seek ways to enhance the reliability and validity of speaking tests.

## II. Literature review

This part presents a number of issues related to the testing of speaking and two important test qualities, validity and reliability, which serves as theoretical framework of the study.

* Faculty of English - Hanoi Open University

### *2.1. Definition of Speaking*

According to Fulcher [6], speaking is the verbal use of language to communicate with others. The purposes for which we wish to communicate with others are so large that they are innumberable. The outwward manifestation of speech is found in sound waves. It's meaning lies in the structure and meaning of all language, whether it is written or spoken. Howerver, speaking difference from written language in a number of respects (Halliday [9]). It is common to note that speaking is usually (although not by any means always) less formal in use of vocabulary, uses fuel full sentences as opposed to phrases, contains repetitions, repairs and has more conjunctions instead of subordination.

### *2.2. The Construct of Speaking*

Fulcher [6] pointed out that there are many factors that could be included in the definition of the construct:

Phonology: the speaker must be able to articulate the words, have an understanding of the phonetic structure of the language at the level of the individual word, have an understanding of intonation, and create the physical sounds that carry meaning.

Fluency and accuracy: these concepts are associated with automaticity of performance and the impact on the ability of the listener to understand. Accuracy refers to the correct use of grammatical rules, structure and vocabulary in speech. Fluency has to do with the 'normal' speed of delivery to mobilise one's language knowledge in the service of communication at relatively normal speed. The quality of speech needs to be judged in terms of the gravity of the errors made or the distance from the target forms or sounds.

Strategic competence: Strategic competence includes both achievement strategies and avoidance strategies. Achievement strategies contain overgeneralization/ morphological creativity, approximation: learners replace an unknown word with one that is more general or they use exemplification, paraphrasing, word coinage, restructuring, cooperative strategies, code switching and non-linguistic strategies (use gestures or mime, or point to objects in the surroundings to help to communicate). Avoidance or reduction strategies consist of formal avoidance and functional avoidance. Strategic competence includes selecting communicative goals and planning and structuring oral production so as to fulfill them.

Textual knowledge: competent oral interaction involves some knowledge of how to manage and structure discourse, for example, through appropriate turn-taking, opening and closing strategies, maintaining coherence in one's contributions and employing appropriate interactional routines such as adjacency pairs.

Pragmatic and sociolinguistic knowledge: effective communication requires appropriateness and the knowledge of the rules of speaking. A range of speech acts, politeness and indirectness can be used to avoid causing offence.

### *2.3. Methods of testing speaking ability*

Followings are some useful and potentially valid formats for testing speaking ability suggested by Weir [19]. These methods of testing speaking ability serves as a theoretical framework for reviewing speaking 6 tests format at Faculty of English, Hanoi Open University.

*2.3.1. Verbal essay*

The candidate is asked to speak (sometimes directly into a tape recorder) for three minutes on either one or more specify general topics. The advantage of this method is that the candidate has to speak at length which enables a wide range of criteria including fluency to be applied to the airport

*2.3.2. Oral presentation*

The candidate is expected to give a short talk on a topic which s/he has either been asked to prepare beforehand or has been informed of shortly before the test. It is different from the 'spoken essay' described above in so far the candidate is allowed to prepare for the test.

*2.3.3. The free interview*

In this type of interview the conversation unfolds in an unstructured fashion and no set of procedures is laid down in advance. Because of its face and content validity in particular the interview is a popular means of testing the oral skills of candidates. Furthermore, interviews are like extended conversations and the direction is allowed to unfold as the interview takes place. However this method is time consuming and difficult to administer if there are a large numbers of candidates.

*2.3.4. The controlled interview*

In this procedure there are normally a set of procedures determined in advance for eliciting performance. With this method, candidates are asked the same questions and thus it is easier to make comparisons across performances. Weir, Cyril J. (1990) stressed that with sufficient training and standardization of examiners to the procedures and scales employed, reasonable reliability figures can be rich with this technique.

*2.3.5. Information transfer: questions on a single picture*

The examiner asks the candidate a number of questions about the content of a picture which s/he has had time to study. The questions maybe extend it to embrace the thoughts and attitudes of people in the picture and to discuss future developments arising out of what is depicted.

*2.3.6. Interaction tasks*

Interaction tasks can be subdivided into information gap between one student and another student and information gap between a student and an examiner. With the first task type, students normally work in pairs and each is given only part of the information necessary for completion of the task. They have to complete the task by getting missing information from each other. Candidates have to communicate to fill in an information gap in a meaningful situation. The second task type can help avoid the possibility of an imbalance in candidates' contributions to the interaction. To examine candidates separately they can be given a diagram, a set of notes, etc. from which information is missing and their task is to request the missing information from the examiner.

*2.3.7. Role-play*

With role-play, the candidate is expected to play one of the roles in an interaction which might be reasonably expected of him or her in the real world. The interaction can take place between two students or between the student and the examiner. However, the disadvantage of the latter is that it is difficult to make an assessment at the same time as taking part in the interaction.

### *2.4. Rating scales*

Rating scale is an interchangeable term with scoring rubric or proficiency scale (Fulcher [7]; Fulcher & Davidson [8]). According to Fulcher [7], "the purpose of the rating scale is to guide the rating process". There are two types of rating scales: holistic and analytic scales.

Holistic scales are on the basis of an overall impression. Raters match students' performance with one of a range of descriptions on scale. Teachers who have enough experience and specialized training tend to select holistic scale (Madsen [15]). But it is not easy to interpret students' scores because each rater has his own criteria in his mind. Furthermore, it does not provide useful feedback for students in order to improve their speaking skills. On the otherhand, analytic scales have been found more reliable than holistic scales even though holistic scales are acceptable. Analytic scales include a number of criteria such as accuracy, fluency, pronunciation, etc., and each criterion has descriptors at the different levels of the scale (Luoma [14]). Raters need to decide how each criterion will be weighted because some criteria may be weighted more heavily, or vice versa. Compared to holistic scales, analytic scales are particularly useful for inexperienced raters to train and standardize them (Weir [20]).

### *2.5. Validity and Reliability*

In order to design a good test, language teachers should understand characteristics of tests. There are some basic characteristics of tests, two of which are absolutely crucial. These two characteristics are validity and reliability. This facet of testing has been discussed at great length by many testing specialists such as Heaton [11,12], Alderson et al. [1], Bachman [2], and Hughes [13].

#### *2.5.1. Validity*

According to the scholars in the field of language testing, the validity of a test is 'the extent to which it measures what is supposed to measure and nothing else'. Validity is an important quality of a test since if a test is not valid for the purpose for which it was designed, the scores do not mean what they are believed to mean. However, Messick [17] defined validity as "an overall evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretation and actions based on test scores". To those familiar to defining validity as "Does the test measures what it is supposed to?", Messick's definition was met with mixed reactions, but in fact Messick had not radically changed the conception of validity held by many researchers in educational measurement. He extended the complex conception of validity that had been evolving for years. Messick's conception of validity can be summarized as follows:

- Validity is not a property of tests themselves; instead, it is the interpretations and uses of tests that can be shown to be more or less valid.

- Validity is the best thought of as one unitary conception, with construct validity as central, rather than as multiple validities such as 'content validity', 'criterion- related validity' or 'face validity'.

- Validity encompasses the relevance and utility, value implications and social consequences of testing. This scope for validity contrasts with the view that validity refers only to technical considerations.

- The complex view of validity means that validation as an ongoing

process of inquiry. The focus on the process of investigation contrasts with a product-oriented perspective of a validated test.

*2.5.2. Reliability*

As previously stated, reliability is one of the crucial characteristics of a good test. Reliability is of primary importance in the use of both public achievement and proficiency tests and classroom tests. Reliability is defined as "the extent to which a test produces consistent results when administered under similar conditions" (Hatch and Farhady [10]). If a test is administered to the same group of students on different occasions ( provided that no language practice work took place in the interval) and if the results are similar the test it is described as reliable.

According to Weir [19]), three aspects of reliability are usually taken into account. The first concerns the consistency of scoring among different markers, e.g. when making a test of written expression. The degree of internal-marker reliability is established by correlating the scores obtained by candidates from marker A with those from marker B. The second concerns the consistency of each individual marker (intra- marker reliability). The third aspect of reliability is that of parallel forms reliability the requirements of which have to be borne in mind when future alternative forms of a test have to be devised. Davies [5] stressed that 'reliability is the first essential for any test but for certain kinds of language test may be very difficult to achieve'.

Heaton [11] pointed out the factors affecting the reliability of a test:

(i) The extent of the sample of material selected for testing: whereas validity is concerned chiefly with the content of the sample, reliability is concerned with the size. The larger the sample (i.e. the more tasks the testees have to perform) the greater the probability that the test as a whole is reliable.

(ii) The administration of the test: is the same test administered two different groups under different conditions or a different times? Clearly there is an important factor in deciding reliability, especially in tests of oral production and listening comprehension:

(iii) Test instructions: are the various tasks expected from the testees made clear do all candidates in the rubrics

(iv) Personal factors such as motivation and illness

(v) Scoring the test.

*2.5.3. Reliability versus Validity*

It is obvious that test validity and reliability are two chief criteria for evaluating any test. However the fundamental problem lies in the conflict between reliability and validity. The ideal test should of course be both reliable and valid. However the greater the reliability of the test, the less validity it usually has. Thus the real life tasks contained in such productive skills tests as the oral interview, role-play, letter writing may have been given high construct validity at the expense of reliability.

## III. The methodology

### *3.1. Methods*

As stated above, the study was designed to ascertain the degree of the reliability and validity of the speaking tests at Faculty of English, Hanoi Open University. Therefore, both descriptive and qualitative research methods were employed in this study to provide a thorough understanding of the assessment

of the speaking tests. According to Creswell & Clark [4], "qualitative data provide a detailed understanding of a problem". However, each of the research methods has its own limitations, and the limitations of one method can be neutralized by the strengths of the other method.

### *3.2. Participants*

Taking part in the study were 8 teachers, three males and five females. Among these teachers, one of them is a PhD and other seven are Masters in Linguistics with TESOL Concentration, who all have over 20 years teaching experience at Faculty of English, Hanoi Open University. These participants have been directly invoved in the teaching and assessment of speaking in general and speaking 6 at English in particular at Faculty of English, Hanoi Open University.

### *3.3. Data collection instruments*

There are several ways to collect data like questionnaire, observation, field notes, interview, documentation, test, and et cetera. In this research, the writer gathered the data through the use of speaking tests to evaluate the reliability and validity of the tests based on the theoretical framework presented in the Literature Review. The researcher collected 10 tests which are currently in use to assess the students' speaking ability at Faculty of English, Hanoi Open University. The speaking tests assess students' use of spoken English. The tests have the same format as follows:

Part 1: the teacher asks a few questions about the students' family, hometown and familiar topics, such as travel, holiday, sport, shopping, interests and so on. This part lasts between two to three minutes.

Part 2: the student is given a topic card. The student has one minute to prepare and is allowed to take notes on a piece of paper before speaking about that specific topic up to two minutes. At the end of this part, the teacher asks one or two general questions on the same topic. This part lasts about four minutes.

Part 3: The student will be asked further questions about the topic in Part 2. These will give them the opportunity to discuss more abstract ideas and issues. This section lasts about three to four minutes.

Additionally, the data was collected through in-depth interview questions on the teachers' beliefs and practices of oral production assessment as follows:

1. How long is the Speaking 6 Test?
2. How many parts/sections are there in the test?
3. What aspects of the speaking test do you like most/ least?
4. Do you think the test is valid enough?
5. What rating scale do you use?
6. What kind of measures do you take to ensure a high level of reliability?
7. Do you have any suggestions on how to improve the procedure of assessment of speaking?

## IV. Findings and discussion

Based on the analysis and evaluation of actual speaking 6 tests and in-depth interviews with the teachers at FOE, the writer collected evidence to ascertain the two important qualilties of the speaking tests as follows:

### *4.1. Validity of the Tests on Speaking 6*

From the data analysis of actual speaking tests and interviews with the

teachers, it was found that the speaking 6 tests are highly valid. All the teachers interviewed stated that the speaking 6 tests have high face validity. Looking at the tests, any student tell effortlessly that the tests are designed to assess speaking skills but not writing or any other skills. As Bachman [2] explains, face validity as for whether the test, on the face of it, truly resembles to test what it is intended to assess, from the learner's perspective.

As far as the structural aspect of validity is concerned, the speaking tests have high validity because they are designed with three tasks, each elicits a different aspect of the speaking skills: responding to interview questions, giving a short presentation, giving opinions, agreeing and disagreeing, making speculations about future, etc. Therefore, it should be possible to show that the scores from the different parts of the test reflect the aspect of the construct it is designed to test. Besides, the content of the tests also has a strong relationship with test construct. It can easily be seen that the test content provides the opportunity for the students to demonstrate their ability on specific function in terms of interaction pattern, task input and student output.

In terms of construct validity, the speaking construct is definable as oral proficiency. The oral proficiency is then reducible to four variables: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy and Pronunciation. However, these variables haven't been broken down further to individualized band descriptors, which is an obstacle to enhancing rater reliability discussed later in this part of the article.

### *4.2. Realiability of the Tests on Speaking 6*

Regarding the reliability of the Speaking 6 tests, the research findings indicate that a high level of reliability is demonstrated within the speaking tests themselves. Eight out of eight teachers interviewed agreed that the test duration is long enough (8-10 minutes) to be both reliable and practical to be administered in the context of achievement speaking tests at the Faculty of English. Besides, four fifths of the participants commented that the speaking tasks are well designed with brief and unambiguous instructions, which cause no difficulty to the students in understanding of what they are required to do in the test. The teachers also added that most of the topics for short talks in part 2 and discussions in part 3 of the tests are in accordance with the syllabus of Speaking 6 currently in use at Faculty of English. However, there is still room for improvement of the scoring of the tests. As Heaton [11] pointed out, one factor affecting the reliability is scoring of the test. stated above, even though the criteria for assessing speaking performance are provided, there is a lack of detailed band descriptors. This is the reason why all of the teachers interviewed said they had to rely on their experience and impression to rate the students' speaking ability. The writer also found that another source of low marker reliability, through interviews with the teachers, originates from the fact that the teachers are not adequately introduced into the assessment process prior to the speaking test. In fact, they are given a teacher's file which provides them with the marking criteria of the test on the test day. Admittedly, marker variability in any subjectively scored test is unavoidable, but there are ways to reduce it and proper training is among them. Two of the teachers asked also confessed that having to play the role of an interlocutor, a rater and comment

writer about the students' strengths and weaknesses in their speaking performance at the same time not only stressed them but also affected their rating process.

Eight out of eight teachers taking part in the interview suggested establishing a framework for making judgements of the students' speaking performance to enhance intra-rater and inter-rater reliability. They added that even though assessment criteria have been introduced, the framework for rating should be designed as scales. The preparation of such a scale, according to McNamara, T. (2000), involves developing level descriptors, that is, describing in words performances that illustrate each level of competence defined on the scale. So, for each aspect assessed, Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy and Pronunciation, the development of a number of separate rating scales is required.

Rater training is another important thing to improve reliability recommended by almost all of the teachers interviewed. Three of them said that the training may take the form of a moderation meeting. At such a meeting, the assessment process is introduced, then some individual teachers are each asked to provide independent ratings for a sample performance. They are then confronted with the differences between the ratings among the teachers. Discrepancies are noted and discussed in detail. The purpose of such a moderation meeting is to bring about broad agreement on the relevant interpretation of the descriptors and rating categories.

## V. Conclusions and recommendations

In this study, the author has used different dimensions to evaluate the validity and reliability of the speaking 6 tests at Faculty of English, Hanoi Open University, including theoretical, descriptive and qualitative. Based on the results demonstrated in the previous part, the following conclusions can be drawn.

First, the speaking tests demonstrate a moderate degree of validity, in light of such features as multiple task types, a strong relationship between test content and test construct, positive feedback from the teachers and comparison with the teaching syllabus.

Second, from the data of in-depth interviews and test materials, it is found that the speaking 6 tests have an acceptable degree of reliability. The test length, test items and test instructions are satisfactory enough to ensure the test reliability. However, the major problem lies in the lack of detailed rating scales based on the test construct. As recommended by language testing scholars and the teachers interiewed, higher scoring reliability of the speaking tests could be achieved by employing analytic scales with detailed band description of each aspect including fluency and coherence, lexical resource, grammatical range and accuracy and pronunciation.

As a result of the study conducted, the following recommendations are made as to the speaking exam and its implementation to improve its reliability and validity:

(i) It would be better if the Faculty held training sessions on standardization before the speaking exam so that all teachers, especially young and inexperienced ones, could benefit from them. The differences between the raters may be reduced in this way as all the teachers can have the opportunity to understand the procedures and the scoring of the exam before the implementation.

(ii) The speaking topics in part 2 of the speaking tests should be considered carefully so that they are in accordance with the theme introduced in the syllabus and of students' common knowledge since it is their speaking ability which is tested, not their world knowledge.

(iii) The scoring reliability would be remarkably enhanced with the employment of analytic scales.

(iii) The process of designing a good test requires a clear understanding of both validity and reliability of the test format. Therefore, the teachers involved in the speaking assessment should be equipped with expertise knowlege about these issues.

***References:***

[1]. Alderson, J.C et al. (1995). *Language test construction and evaluation*. Cambridge University Press.

[2]. Bachman, L. F. (1990). *Fundamental considerations in Language Testing*. Oxford University Press.

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

[4]. Creswell, J.W. & Clark, V.L.P. (2O11). *Designing and conducting mixed methods research* (*second edition*). California, America: Sage Publications.

[5]. Davies, A. (1990). *Principles of Language Testing*. John Wiley and Sons Ltd.

[6]. Fulcher, G. (2003). *Testing second language speaking.* London: Pearson Education.

[7]. Fulcher, G. (Ed.) (2014). *Testing second language speaking (Second Edition).* London & New York: Routledge.

[8]. Fulcher, G. & Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. New York: Routledge.

[9]. Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.

[10]. Hatch,E. & Farhady, H. (1982). *Research design and statistics for applied linguistics*. Newbury House Publishers, Inc.

[11]. Heaton, J.B. (1988). *Writing English Language Tests (new edition).* Longman.

[12]. Heaton, J.B. (1990). *Classroom testing*. Longman.

[13]. Hughes, A. (1989). *Testing for Language Teachers*. Cambridge University Press.

[14]. Luoma, S. (Ed.) (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.

[15]. Madsen, H.S. (1983). *Techniques in testing*. Oxford: Oxford University Press.

[16]. McNamara, T. (2000). *Language Tes*ting. Oxford University Press.

[17]. Messick, S.A. (1989). *Validity*. In R. L. Linn (ed.), Educational Measurement, 3rd edn. New York, NY: Macmillan Publishing Co.

[18]. Nakamura, Y. (1993). *Measurement of Japanese college students' English Speaking ability in a classroom setting.* Unpublished doctoral dissertation, International Christian University, Tokyo.

[19]. Weir, C. J. (1990). *Communicative Language Testing*. Prentice Hall International (UK) Ltd, Great Britain.

[20]. Weir, C.J. (2005). *Language testing and validation*. New York: Palgrave Macmillan.

***Author address: Faculty of English - Hanoi Open University***

***Email: maihuong74@hou.edu.vn***