



Direct Spoken English Testing Is still a Real Challenge to be Worth Bothering About

Majid N. Al-Amri

English Language Center, Yanbu Industrial College

Yanbu, Saudi Arabia

Tel: 96-65-0691-9131 E-mail: majid_yic@yahoo.com

Abstract

This paper introduces and discusses issues related to the challenge of obtaining more valid and reliable assessment and positive backwash of direct spoken English language performance of students in real-life situations. For this purpose, the paper is divided into four sections. The first section is the introduction to the article. The second part includes presentation and discussion of validity, reliability and practicality as principles of direct spoken English tests. Part three consists of an evaluation of the challenging nature of direct spoken English testing for examiners to deal with. The last section is the conclusion of the paper.

Keywords: Direct Spoken English Tests, Nature of direct Spoken English Testing, Challenge in Spoken English Testing, Subjectivity and Time Consuming Nature of Spoken English Testing

1. Introduction

An English language test is necessary and valuable not only for a particular aspect of language skill, but for all of them-including speaking English in real life situations. Speaking English in real-life situations is an important skill that should take priority in any language test. To test this ability, teachers need to employ direct tests in the real life contexts so that they can obtain an idea of what candidates would do in real life situations. However, obtaining more valid, reliable and practical measurements of direct spoken English language performance and more positive backwash is still a real challenge for examiners to be worth bothering about since the nature of language testing is connected to social relationships, power, and control (Shohamy, 2001; Norton & Toohey, 2004, Kang, 2008).

Testing spoken English directly is still a real challenge, especially when it involves different criteria for assessment that may lead to disagreement between testers themselves, e.g. whether fluency or accuracy is being judged. There may also be difficulty in judging whether only speaking or speaking and listening together. In addition, assessors may face a challenge of how to offer an opportunity for students to “demonstrate their knowledge or skills in the content being assessed.” (Young, 2008, p. 2)

Moreover, it is still absolutely impossible to avoid some degree of subjectivity, power and control in assessment regarding, for example, scoring the assessment scale. On the other hand, direct testing has the particular problem of needing the necessary investment of time and money in order to test large numbers of students. However, even when the computer is employed in testing, it is still absolutely impossible to avoid some degree of subjectivity in assessment regarding, for example, scoring the scale and evaluating the importance of items in the part of the course or in real life situations.

2. Principles for Spoken English Testing

Validity, reliability and practicality are important principles for defining a spoken English test. Validity can be defined as the extent to which a test measures what it is supposed to measure. Spoken tests (assessments of English language proficiency) are valid when they are able to give a clear idea about the ability of the candidates to communicate, for example, in real life situations. However, regardless of students’ level of English proficiency, “all students should have an opportunity to demonstrate their knowledge or skills in the content being assessed. Although students may have different English proficiency classifications, the meaning of their scores on content assessments should be comparable.” (Young, 2008, p. 2)

The more features of the real life activity, the easier it will be to translate the performance of the candidate. The face-to-face interview, for example, can help the examiners to determine whether the candidates can speak in real life situations or not. However, direct test formats are usually too time consuming and difficult to administer, especially if there is a large number of candidates. (Weir, 1988) In addition, in general, direct tests are subjective and depend on whether the assessors are good at using grading rubrics/scales and test formats to minimize and reduce subjectivity and disagreement.

Regarding reliability, there are two types. One type is concerned with the examiners. Different examiners are expected to give similar and comparable marks to the same test on two different occasions. (Berkowitz, Wolkowitz, Fitch, and Kopriva, 2000) The second type of reliability is test-retest reliability. It is the ability of a spoken language test to achieve the same result time after time. (Rudner, 2001) Overall, it can be said that a spoken test is a reliable test when it achieves the same result time after time and when its examiners give similar and comparable marks to the same test on different occasions. So using criteria and training in the use of grading rubrics/scales and test formats to minimize and reduce subjectivity and disagreement are highly important to achieve more reliable results.

Like validity and reliability, practicality in testing spoken language is important. It is about, for example, the amount of time that will be spent on a test as well as the availability of equipment and resources which will be of help for examiners. However, existing resources should not be used beyond its limits. (Weir, 1994) This means that the necessary information about the candidates needs to be obtained in a short time. The more direct the test is such as the face-to-face interview, the more time-consuming and less practical to give a valid assessment. So using criteria and training in the use of grading rubrics/scales, test formats and equipment and resources available are very important with regards to increasing practicality of a direct English language test.

It can be noticed that there is a relationship between validity, reliability and practicality. For example, if a spoken test is valid, it must be reliable as well. In other words, if a spoken test is able to test what the examiner wants it to test, it must also be able to give the same result time after time. When a test gives different results at different times it cannot be valid. However, it is possible for a spoken test to give the same result at different times without being valid. In addition, validity may be threatened by practicality. This means that making direct tests as practical as possible may lead to shorter test time and to them being less direct. Again, using criteria and training in the use of grading rubrics/scales and real-life test formats in their contexts can increase not only validity of the test but also its reliability and practicality.

3. Direct Spoken English Testing as a Challenge for Examiners

Direct spoken English testing is a challenge in nearly all aspects, e.g. in the design of tests, the producing of test items, the determining of test scores and the setting of time limits and other administrative procedures. (Pilliner, 1968; Bachman, 1990; Ur, 1996) All of these sources of the challenging nature of direct spoken English testing can affect both reliability and validity of test results. In the following, different sources of challenging will be presented and discussed.

3.1 Time

Examiners still face the challenge that each candidate has to be tested in real life time. The oral test should be made as long as is feasible. Sometimes it is impossible to obtain a reliable test in less than 15 minutes, but 30 minutes is quite enough for obtaining all the information necessary for most purposes. (Hughes, 2002) However, time needed depends on the level and the purpose of the test. A time limit of 10-15 minutes is normal in most short/indirect published tests. One issue would be: is it worth for the sake of backwash having a shorter test, or an indirect test?

Direct testing of spoken English is time-consuming and subjective in all its stages. (Hughes, 2002) At the preparation stage, for example, a considerable time is required for the sharing of work as well as the use of mechanical tasks such as checking answer keys. At the test operation stage, it is a challenge for examiners to decide how much time will be spent on carrying out the test procedures and the amount of time that students have to spend on testing. At the final stage, the test improvement stage, making adjustments to the techniques, making systems and monitoring the test all might affect the amount of time that can be spent on testing.

3.2 Test formats

There is a range of formats of varying degrees of directness. It includes the more direct type (e.g. the face to face interview) and the more indirect multiple choice pencil and paper test which can be scored by computer. Also, test formats can be classified into three groups formats: indirect formats (e.g. sentence repetition, the mini-situation tape, information transfer); interaction student with student formats (e.g. information gap exercises); and interaction student with examiner or interlocutor formats (e.g. the free interview/conversation, the controlled interview, role play and information gap). (Weir, 1988, 1994) Each group has its own advantages and disadvantages, and the teacher/assessor should decide which would be most appropriate for the students. In other words, there are still important questions remaining to be answered by examiners such as what principles should underlie good test formats, what type of test formats should be used for large numbers, is it worth having a shorter test format, or an indirect test format, etc.

3.3 Real life routines

Keeping interaction in the routines is very important. This means that test formats should include certain routines which reflect real life situations such as expository routines (e.g. description and narration) or evaluative routines (e.g. preferences or the drawing of conclusions). Furthermore, using various improvisation skills, such as checking understanding or expressing agreement or disagreement, plays an important role in interaction, helping, for example, to check specific information or correct mistaken interpretations. (Bygate, 1987) However, in some situations, there are

thousands of candidates, which means it is an immense task even if only a few minutes is given to test each candidate. This also demands sufficient resources. For example, it may prove very difficult to record thousands of candidates as well as to employ the proper numbers of people to administer the test. It is actually not practical and it demands a large amount of time.

3.4 *Psycho-sociolinguistic relationships*

The practice of direct language testing was expanded to include the importance of context beyond the sentences to appropriate language use. Accordingly, the relationships between the examiner(s) and the candidate(s) are psycho-sociolinguistic relationships connected to power and control. (Shohamy, 2001; Norton & Toohey, 2005; Kang, 2008) For example, the tests which are administered by familiar personnel are better than those administered by unfamiliar personnel. The candidates will also find it easy to speak to people with whom they are familiar or similar in status, e.g. speaking to a single peer is easier than speaking to an unknown authority figure. (Bachman, 1990; Weir, 1994) In addition, the number of examiners who usually have different questions and ways of communication and understanding can affect the test taker's responses. (Shohamy, 1982; Scott 1986; Bachman; 1990) Also, differences in sex have an impact on the test taker's performance. Women are usually better than men at keeping a conversation going. (Weir, 1994) Furthermore, "speech assessment is sensitive to rater's expectations and social stereotypes." (Kang, 2008, p. 201)

3.5 *Reciprocity*

Reciprocity is a feature of speech which crucially affects the decisions of the examiners that have to be made. (Bygate, 1987) Examiners have to pay attention to their examinees as well as adapting their messages according to their reaction and the type of speech (monologue, conversation and interview). For instance, the candidate will find it easier to carry out a face-to-face conversation than a telephone conversation with an interlocutor in a different room. In addition, the candidate's performance can be affected by features of the language used by the interlocutor, e.g. the rate of utterance, the accent of examiner, the clarity of articulation of the examiner and the length of discourse. Also, each individual factor of raters' characteristics can affect the rating of oral assessments. For example, raters from different language backgrounds (NS/NNS, exposure to NNSs' contact, and prior teaching experience) often have different perceptions on the international testees' accented speech (speech rate, pauses, and stress), and the intercultural intervention (rater training) usually exerts an impact on accentedness rating. (Kang, 2008)

There is some evidence supporting the use of the computer in assessing spoken language performance. Jared points out that using the computer may replace traditional spoken language testing methods. He states that analysis of data, from over 10,000 subjects collected as a result of a series of experiments carried out at 18 colleges and universities in the U.S., Japan and Italy, provides evidence in support of the validity (independent of the learner's other social and cognitive skills) as well as the reliability (split-half correlation of 0.94) of the computer-based scores.

However, in spite of the advantages of computer multimedia, computer programs are still imperfect and they need to be intelligent enough to interact with learners and understand their way of thinking and behaving like their teachers might do. For example, they mainly deal with reading, writing and listening and "even though some speaking programs have been developed recently, their functions are still limited." (Lai & Kritsonis, 2006, p. 4) They need to be able to understand learner's spoken input and evaluate it not only for correctness but also appropriateness (Warschauer, 1996). This shows a fundamental difference in the way humans and computers utilize information (Blin, 1999; Dent, 2001).

3.6 *The optional cut-off technique*

Using the optional cut-off technique which probably leads to saving much time is recommended in implementing direct testing. (Underhill, 1987) It can be used when the learners can be seen not to be taking the same amount of time. In other words, the interviewer can use graded tasks in order of increasing difficulty. When the candidate has clearly failed two or three successive tasks, the interviewer can confirm his/her level by returning to a slightly lower level for another two or three items before ending the test. This technique can be used with several print formats of test such as matching appropriate responses, question and answer and using a picture. However, it is still impossible for examiners to avoid some degree of subjectivity in the use of the cut-off technique, especially with a large number of candidates who need to be tested in a limited time.

Recently, digital formats like BEST Plus assessment formats have been designed to be delivered via computer. (CAL, 2009) The test examiner asks the examinee a question presented on the computer screen, listens to the examinee's response, uses the scoring Rubric to determine the scores for that item, and enters the scores into the computer. The computer then selects the next test item, choosing items most appropriate for the examinee according to the scores entered for previous responses. The examinee might not see the same test twice. However, just like with print formats, it is still impossible to avoid some degree of subjectivity regarding, for example, how much time should be spent on each item of the test, how important is this element in the part of the course or in real life situations, how essential the candidates should know why they are doing the task.

3.7 Environmental conditions

Test tasks might be expected to be performed differently under different environmental conditions. (Bachman, 1990) So the test tasks should reflect the target real life situation. "If the candidates are placed in a setting, say for a role play which is not likely to reflect their future language-using situations, validity is impaired to that extent." Weir (1994, p. 38) Furthermore, testing spoken language should be carried out in a quiet room with acoustics. (Hughes, 2002) However, full authenticity of setting is obviously not attainable but settings should be made as realistic as possible (e.g. by creating imagination on the basis of carefully designed rubrics). (Weir, 1994)

3.8 Certain arrangements

Certain arrangements can facilitate testing. For example, the furniture in rooms used for testing can be arranged in a way that facilitates testing, e.g. the chairs of the interviewers can be arranged at an angle or even placed side by side so that testing takes place easily. (Underhill, 1987) In addition, using available resources can also help to facilitate test, such as using reliable equipment which is easy to use for the test requirements. However, although organizing settings and using reliable equipment can facilitate testing, there are still challenging questions remaining to be answered, especially with testing huge numbers of candidates such as is it their responsibility to make certain arrangement or someone else, are examiners aware of the importance of such certain arrangements, do examiners have time to make certain arrangements etc.

3.9 Test scales

There are two types of scales for testing spoken language directly. They are the holistic scale and the analytic scale. The holistic scale is a general scale for overall speaking ability. The analytic scale is a detailed scale for several aspects of the skill of speaking such as grammar, pronunciation, vocabulary etc. However, obtaining valid and reliable scoring is still one of the major difficulties in testing speaking. Examiners still face the challenge to develop a scale that can be applied as objectively as possible. Although the computer can be used for scoring direct testing, it seems that it is still absolutely impossible to avoid some degree of subjectivity in assessment regarding scoring the scale and evaluating the importance of items in the part of the course or in real life situations.

However, each one of the analytic and holistic scales has its own advantages and disadvantages. The choice between them usually depends on the purpose of the testing. The holistic scale is more economical with time when it is carried out by a small, well-knit group at a single site. (Hughes, 2002) The analytic scale is more appropriate if it is to be conducted by a heterogeneous, possibly less well trained group, or in a number of different places such as in the IELTS test. In addition, the analytic scale is essential when diagnostic information is required. Although they can be used together and they can be conducted by more than one person so that their reliability can be checked, examiners still need to be well-trained and familiar with the structures of a test, such as the tasks, the timing and the scoring sheets etc. They also need to be aware of any irrelevant feature of performance. (Hughes, 2002) Furthermore, it is still a challenge for examiners to record the speaking so that the scoring can be done from the tape, especially when they are huge numbers of examinees.

4. Conclusion

Obtaining more valid and reliable measurements of direct spoken language performance (language proficiency) and more positive backwash is still a challenge for examiners. It requires considerable time, efforts and familiarity with students' experiences, knowledge and skills in the content being assessed. Assessors/teachers who make a serious commitment to using criteria, training in the use of scales, and identifying and reflecting on their previous experiences, power, control and subjective decisions can plausibly anticipate more valid, reliable and practical assessments. This would also be a great opportunity to invite students to become self-assessors themselves. This cooperative practice of testing/assessment would enhance student motivation, confidence, and achievement.

References

- Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Berkowitz, D., Wolkowitz, B., Fitch, R., Kopriva, R. (2000). *The Use of Tests as Part of High-Stakes Decision-Making for Students: A Resources Guide for Educators and Policy-Makers*. Washington, DC: U.S. Department of Education. [Available online: <http://www.ed.gov/offices/OCR/testing/>].
- Blin, F. (1999). CALL and the development of learner autonomy. In R. Debski & M. Levy (Eds.). *World CALL: Global perspectives on computer-assisted language learning* (pp. 133-147). Lisse: Swets and Zeitlinger.
- Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.
- CAL (2009). *Testing / Assessment*. Retrieved August 10, 2009 from, <https://www.cal.org/>.
- Dent, C. (2001). *Studer: classification v. categorization*. Retrieved May 23, 2006, From <http://www.burningchrome.com:8000/cdent/fiaarts/docs/1005018884:23962.html>.

- Hughes, A. (2002). *Testing for Language Teachers*. Cambridge: Cambridge University Press.
- Jared, B. Computer evaluation of spoken language performance. Retrieved August 10, 2009 from, http://www.languages.dk/eurocall/eurocall00/computer_evaluation_of_spoken_la.htm.
- Kang, O. (2008). Ratings of L2 Oral Performance in English: Relative Impact of Rater Characteristics and Acoustic Measures of Accentedness. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 6, 181–205.
- Lai, C. C., & Kritsonis, W. A. (2006). The advantages and disadvantages of computer technology in second language acquisition. Doctoral Forum: *National Journal for Publishing and Mentoring Doctoral Student Research*, 3(1), 1-6.
- Norton, B. & Toohey, K. (eds.) (2004). *Critical pedagogies and language learning*. Cambridge: Cambridge University Press.
- Pilliner, A.E.G. (1968). Subjective and Objective Testing. In A. Davies (ed.). *Language testing symposium: A psycholinguistic approach*. (pp. 19-35). Oxford University Press.
- Rudner, L. (2001). *Reliability*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
- Scott, M. L. (1986). Student affective reactions to oral language tests. *Language Testing*, 3, 99-118.
- Shohamy, E. (1982). Affective considerations in language testing. *The Modern Language Journal*, 66, 13-17.
- Shohamy, E. 2001. Democratic assessment as an alternative. *Language Testing* vol. 18, no. 4, 373-391.
- Underhill, N. (1987). *Testing Spoken Language*. Cambridge: Cambridge University Press.
- Ur, P. (1996). *A course in Language Teaching*. Cambridge: Cambridge University Press.
- Warschauer, M. (1996). Computer assisted learning: An introduction. In S. Fotos (Ed.), *Multimedia language teaching* (pp. 3-20). Tokyo: Logos International.
- Weir, C. (1988). *Communicative Language Testing with Special Reference to English as a Foreign Language*. University of Exeter Press.
- Weir, C. (1994). *Understanding and Developing Language Tests*. London: Prentice Hall.
- Young, J. W. (2008). Ensuring valid test content tests for English language learners. R&D Connections 8. Princeton, NJ: Educational Testing Service.