An Analysis of Speaking Test Design for the Undergraduate Students in University Y in Guangdong Province, China

Wei Zhou

Anhui Vocational And Technical College, Hefei, Anhui, 230011, China

Abstract: Speaking competence is of great importance in using English as a global language. However, there exists a shortage of speaking assessment for undergraduate students in our country. To address the gap, this study designs and analyzes a speaking test for undergraduate students in University Y, China. In order to design a valid, reliable speaking test, the principle of authenticity, practicality, washback, motivation have been taken into consideration.

Keywords: Communicative competence; Speaking assessment; Undergraduate students; Validity; Reliability; Authenticity; Practicality; Washback; Motivation

DOI: 10.12346/fhe.v3i3.3987

1. Introduction

Globalization has brought about an increasing need of intercultural communication, political negotiation, business cooperation and economic development, all of which call upon the communicative competence to use English as a global language [1]. In response to this, in China, English curriculum embraced a reform at national level. The speaking skill is increasingly valued.

In reality, the situation seems different. As known to all, EFL testing is of great importance in Chinese education. English is a compulsory subject at all levels of education in China and therefore playing an important role in different high-stakes exams. For example, in higher education of China, the College English Test (Band 4) is a national norm-reference test, without passing which students cannot get a bachelor degree [2,3]. Surprisingly, in such an important test, speaking skills are not assessed due to the limitation of resources. Moreover, in most of EFL tests in China, speaking tends to be ignored by both teachers and students as well. There seems to be a shortage of speaking tests and passion to design speaking tests for undergraduate students in China. Therefore, in this essay, I am going to design and analyze a speaking test for undergraduate students in University Y, China. This essay starts with an introduction, and then comes a context section. Section three is a discussion of assessing speaking. Section four will focus on the test design and its rationale. Concluding this essay is a brief summary and implication.

2. Context

2.1 Describing the Teaching Context

The teaching context is University Y in Guangdong province, China. It is this university where I have been studying for four years. Therefore, I choose the EFL speaking course of senior students (non-English major) in university Y as the context of this essay.

In university Y, the oral English class size is approximately 20 students. The teachers for oral English course are native speakers of English language from Britain, America, Australia, and Canada. The teaching material used is a series of textbooks called *Inside Out*, published by Shanghai Foreign Languages Education Press in April 2007. However, in reality, the foreign teachers often put the textbooks aside, and design their own oral English class. Although specific course content differs from teacher to teacher, the entire oral English courses in university Y share the same course objectives claimed by the College English teaching syllabus.

According to The College English teaching syllabus (2010), oral English course objectives of senior students in university Y includes “At the end of the course, students can conduct a conversation in relation to daily life topics; students can give a brief presentation of common topics in social life; students can express themselves coherently and fluently; students can produce English in a natural intonation and with accurate pronunciation; students should achieve an accuracy of grammar and appropriate use of English overall”.

As for the assessment of oral English course, there are two formal tests in each term: mid-term tests and final-term tests and the test results are recorded in the university system. In terms of task type, it is usually role-play for a certain situation in given time between two examinees.

2.2 Test-takers

The target students are senior students aged from 21 to 24. Most of them are upper intermediate level proficiency, and have passed CET4. However, in reality, due to the limitation of the resources, the speaking part of CET4 is inaccessible to students. Therefore, CET4 tends to have a negative influence on developing students' communicative competence, for students are not assessed for speaking [4]. In addition, as discussed above, although the mid-term and final-term speaking tests of senior students in University Y have communicative activities such as role-play, that is the only task type involved in the speaking test.

The target students in University Y have showed a limited ability in speaking English, and the teaching style they got used to is grammar-translation approach, however, they gradually realized the significance of communicative competence and claimed their needs of participating in more communicative activities in EFL classroom, especially in oral English course.

3. Assessing Speaking

3.1 Significance of Speaking

Among the four language skills (listening, speaking, reading and writing), speaking is likely to be the most important one [5-7]. The evidence might be found from the use of 'speaker of a language' that refers to those people who know and use that language very well, such as "native speaker', or 'English-speaking countries'. In this sense, it seems that speaking involves all the other language skills. That is to say, the ultimate goal of language learner may be able to speak the target language fluently [8].

As a global language in the world, English speaking proficiency is of key importance to China's involvement in the diversified and vivid international communication. As a consequence, communicative language approach (CLT) was added to the national English language curriculum of China. Communicative competence has been valued in language learning and teaching [9]. Speaking is considered as a significant element of CLT in EFL classroom. Moreover, the view of "learn to communicate by communicating" also suggested the importance of speaking [10].

3.2 Significance of Assessing Speaking

From the perspective of learners, speaking tests are complex and challenging for they have to learn the language knowledge and also the skills to use that knowledge. For teachers or examiners, speaking tends to be one of the most difficult skills to be tested as well. One of the reasons may be that subjectivity is hardly to be avoided in the process of assessing.

Although the difficulty and complexity of speaking tests existed, still, there are reasons to assess speaking in EFL learning and teaching. As discussed in the preceding section, speaking plays a crucial role in language learning and teaching. Therefore, in order to improve students' speaking ability, their progress of learning could be tested and analyzed. In terms of washback, the existence of speaking tests may encourage more devotion of learning and teaching of speaking in and outside of EFL classroom as well [11]. In addition, speaking tests have experienced dramatic change over the recent decades. For example, the focus of spoken grammar and pronunciation in speaking test has shifted to genuine communication [12].

4. Test Design and Rationale

4.1 Purpose of the Test

The test to be designed for the oral English course of senior students in University Y is an achievement test and a criteria-referenced test (CRT). Achievement test is "to measure learners' ability within a classroom lesson, unit or even total curriculum" [13]. The primary aim of an achievement test is to examine whether students have achieved the learning objectives at the end of the course. The speaking test I am going to design is mainly for assessing students' learning outcomes of oral English course at the end of university year. This test is summative, and it will be administered in the final week of the academic term. The grades of this speaking test will be recorded in students' profile in university system. Even though it is a summative assessment, the test can also serve a formative function. For example, examinees' performance in different subsets of the course could be offered feedback. Furthermore, it is CRT test, in this sense, students' performance in the speaking test will not be compared to other students' or any 'norm', but only to test the amount of knowledge and skills learners required in relation to the course objectives.

4.2 Construct Specifications

4.2.1 Concept of Construct

"Testing the ability to speak a foreign language is perhaps the least developed and the least practiced in the language testing field." (1961). The main reason is possibly a lack of understanding towards the construction of speaking. However, it seems difficult to give an operational definition for the construction of speaking. For instance, Butler et al. [14] have tried to define the construct of speaking in academic settings, but ending up describing speech from the perspective of sociological and speech act theory rather than offering a definition.

Regarding this issue, the solution provided by Lado (1960) is to include mere linguistic elements in the definition, for the purpose of clarifying the concept of speaking without any other variants, such as "talkativeness" and "introversion". However, a more recent approach tends to cover contextual factors in construct definition, and especially in the situation of English for specific purposes (ESP) [15]. Besides, some other factors could also be incorporated into the concept of speaking construct. For instance, since interaction occurs between participants, then the degree of the support and cooperation of interlocutors may also be taken into consideration.

From the discussion above, it can be seen that the consensus on the definition of construct is difficult to achieve. Specifically, it is very challenging to clarify the dynamic, complex human communication in any operational definition. Furthermore, it is also unnecessary to cover each component of speaking construct in any particular test. Therefore, the choice of how to define construct mainly depends on test purpose and also to what degree the scoring could be generalized in different contexts [16].

4.2.2 Construct to be Tested in this Context

Considering the test purpose in my context, the construct of speaking is mainly based on the checklist designed by Bygate (1987). According to this checklist, the construct of speaking includes three sub-skills: Routine skills (information routines, interaction routines), improvisation skills (negation of meaning, management of interaction), and micro-linguistic skills. In order to fit the test purpose stated in the previous section, 'interaction routines' and 'management of interaction' skills will be excluded in my test design.

The routine skills to be tested here mainly focus on "information routines", which refers to "conventional ways of presenting information" and "frequently recurring ways of structuring speech", such as narration, describing, comparison and so on [17]. It may involve identifying and sequencing some subjects or explaining, reasoning and evaluating. When the communication based on routine skills break down, the improvisation skills can function to continue the interaction. The negotiation of meaning could be defined as the efforts to maintain the interaction by adjusting the conversation or clarifying the misunderstanding (1993), such as checking understanding, responding, establishing the friendly atmosphere. As for micro-linguistic skills, it includes the accuracy of grammar, the intelligibility, the variety of vocabulary use and structure use (1993).

4.3 Task Specifications

4.3.1 Task Types and the Target Skills to be Tested

Speaking tasks refer to those activities involving speakers to use the target language for achieving specific goals or objectives in certain speaking contexts. The task types chosen in my test design are narrative task and controlled interview task.

The narrative task

The rationale:

For this task, the examinee will be given a piece of paper, on which presents a series of pictures in chronological order (from year 2000 to 2008) and they are required to narrate a story based on these pictures. (See appendix1). The primary purpose to design this narrative task is to test students' routine skills in speaking. Generally, it provides students a sequence of events to describe and narrate. To be specific, firstly, coherence, organization of describing and narrating can be assessed from the long turn of students. Secondly, the time framework is set in past tense; therefore, grammar knowledge of past tense could be tested in this task as well. Lastly, the features of narration are tested, such as setting the scene, identification, and coherent description.

Advantages:

One obvious advantage may be that the task requirement is very clear. This task only deals with pictures, rather than reading or listening materials (1993). In this case, it avoids the influence of other measurement. Moreover, for examinees, reading pictures tends to be quicker than written materials (if the pictures are good-designed), thus saving precious and expensive exam time.

Limitation:

The potential danger in this narrative task is the ambiguity about the pictures. For example, some culture-loaded messages in pictures may pose challenge to examinees, besides, the use of maps or abstract pictures may pose cognitive complexity. However, the test aims to assess speaking ability, not the ability to read maps or anything else. The physical condition of the pictures may be another problem [18]. All the potential problems here will be discussed with suggestions to cope with them in validity and reliability section.

The controlled interview task (see Appendix 1)

The rationale and advantage:

Interview is among one of the most common used tasks in assessing speaking. The structured interview designed in this test is to assess improvisation skills. For instance, the examinees may ask for repetition, and examiners could ask for clarification of responses. As for its advantage, the controlled interview is designed with some standardized structured questions. This is comparatively easier for training interlocutors and raters. In this sense, it may also improve validity and reliability of speaking test.

Limitation

The most obvious potential problem is possibly the subjectivity in the interview. For instance, even there are standardized question types, examinees may be given different topics and examiners perhaps ask questions in different manner, which may cause unfairness to some extent (1990). In addition, students may feel nervous communicating with examiners.

4.3.2 Instructions for Test-takers and Administration Plan

This speaking test consists two parts: narrative task and controlled interview. The instructions of tasks will be presented in written form in a piece of paper that will be given to students. However, the examiner will still claim the instructions in the beginning of the test. The principle of designing instructions language is as simple and guided as possible, for the supporting materials should function as a support rather than a burden to examinees (2004).

The goal of administration plan is to ensure all the participants (examiners, examinees) know what they should do at what time. All the resources and arrangements in relation to the test should be checked before tests start. For example, in my context, the room for interview should be quiet, and audio recorder is in good condition. The examiners should arrive around 30 minutes before exam time. Moreover, the examiners are given rater training for some time.

4.4 Assessment Specifications

The marking criterion to be used in this test is Analytic-marking scheme (see appendix) designed by Weir (1993). This criterion covers six parts: appropriateness, adequacy of vocabulary, grammatical accuracy, intelligibility, fluency, relevance and adequacy of content and each part is scored into four levels (1993). The main reason to use this analytical rating scale is to provide comparatively detailed feedback to students. At the same time, it can also serve the function of a diagnostic test showing leaners' strength and weakness of speaking, which may be beneficial to future course design as well. The details of how to use this rating scale will be discussed in validity section.

4.5 A Reflection of Principles of Testing

4.5.1 Validity (including washback)

The concept of validity

Validity is of primary importance in test, for a test is meaningless without validity. Validity is traditionally defined as 'measures what it is supposed to measure' (2012). A new recent approach to investigate validity is suggested by Messick (1996). This approach focuses on six aspects of construct validity: content aspect, substantive aspect, structural aspect, generalizability, external aspect, and consequential aspect. For my test, I will analyze the content aspect, structural aspect and consequential aspect of validity.

Content aspect of validity concerns whether the test content is adequate, relevant, and appropriate from the perspective of difficulty level and items design. Moreover, structural aspect deals with the question of whether the scoring rubrics fit the construct behind the test. As for consequential aspect validity, it is mainly about whether the test is fair, non-biased and also the intended or unintended consequence of tests, including washback.

A reflection of validity in my test

As for the content aspect of validity, the test is a direct test that aims to assess speaking, for example, in the narration task; it tests speaking skill directly without any other variance of measurement. The difficulty of task items is appropriate to senior students in University Y; for example, the narration task is comparatively easier, and then followed by structured interview. In addition, two experts in test deign will be invited to evaluate the content validity of my test as well. Therefore, I think the content presents a good representative of speaking tests. When it comes to structural aspect, my test indicates a good validity. For instance, my test includes two tasks, which could test the routine skills and improvisational skills respectively. Moreover, the rating scale in this test also indicates exactly the construct to be tested. For example, appropriateness, fluency is designed for testing routine skills and improvisational skills as well, grammatical accuracy, intelligibility, the adequacy of vocabulary and structure is for micro-linguistic skills. Lastly, when it comes to the consequential aspect, let me take washback as an example. The test is designed to test the skills encouraged to use. For instance it focuses on routine skills, improvisational skills and micro-linguistic skills, therefore, both teacher and students are likely to devote more time in these aspects of speaking before and after test.

4.5.2 Reliability

The concept of reliability

Reliability refers to 'the consistency of the scores' (2004). Reliability is very important, in that decisions can only be made upon reliable scores. Otherwise, unreliable scores will cause unfairness in placements, promotions and so on.

A reflection of reliability in my test

Different techniques have been used in the test to ensure reliability. Firstly, rater training is provided. It is almost the most traditional technique to ensure reliability. However, rater training has been criticized for training novice teachers to be doctrinal. Even this may be true; Luoma explained that the evidence to support the validity of criteria could argue against it. Secondly, benchmark tape will be included in rater training. Through comparing and analyzing the previous examinees' performance, raters can know the concrete features of performances at different levels. Besides, two experts are also invited to introduce and analyze the strong and weak performance of samplings respectively (2004). Thirdly, intra-rater reliability could be achieved by asking raters to score the same sample of performance over a period of time, if the raters agree with themselves. Additionally, comparing different raters' rating the same piece of work may enhance inter-rater reliability. Another point to add here is the use of rating form, if two raters disagree with each other on the rating of the same performance, the rating from could serve a good role here, for the detailed notes in rating form could be compared and discussed then. Moreover, as we know that subjectivity inevitably exists in assessing speaking, but the second rater will rate the performance recorded in the tape, and the performance is anonymous. Finally, the second task of structured interview in my test could increase the reliability, because the question types are controlled, and topics are chosen at the similar level of complexity.

4.5.3 Authenticity and Practicality

Authenticity is defined as "the degree of correspondence of the characteristics of a given language test task to the features of a target language task"[19]. Therefore, the more likely to occur in real life, the more authentic the task is (2010). I tend to believe the tasks in my test are authentic to some extent. For example, in the narrative task, examinees are required to describe a story according to the pictures. Telling stories in social life is very normal, especially when chatting with friends. The stories in the pictures are also set in a certain context. Therefore, in this sense, the narrative task in the test is authentic for it serves a communicative purpose and it is meaning-focused. Besides, in the interview task, the topics are very common to social life of university students, such as studying, travelling, friendship, shopping, experience and so on. These topics are meaningful, interesting and relevant to examinee's life as well.

Practicality concerns the resources and facilities in relation to the planning, administration and rating procedures of an assessment. Regarding my test, in the administration stage, first of all, the acceptable environment condition has been ensured. The air conditioner, the audio recorder, the available classroom for the interview is all considered here. When students are doing tasks, if it is expected to be a longer turn, more time for preparation will be provided. Otherwise it is a short period of time. In terms of rating procedure, the issue of ease of rating and ease of analyzing rating has been taken into account. In my context, the speaking class size is 20 students. Therefore, it is practical for me to choose analytical scheme. If it is a big class size, maybe holistic approach is more appropriate due to the limitation of examiners' valuable time (2012). The other reason I choose this scheme is that the rating scales are designed at only four levels, thus not too complex and complicated to rate from the aspect of teachers. Still in this way, students can get concrete but concise feedback from this test. The last point to be evaluated is the contradiction may arise between the ease of scoring and the ease of interpreting the score, a balance is suggested to strike between them.

5. Conclusions

In this essay, a speaking test design for senior students in University Y of China has been analyzed and reflected. Through the reflection, it can be concluded that due to the complex, dynamic and communicative nature of speaking, speaking testing is a challenging field to explore. In order to design a valid, reliable speaking test, the principle of authenticity, practicality, washback, motivation should be taken into consideration. However, the most important point may be to consider the test purpose first, and I think that may be the starting point but also the primary principle among all. One of the limitations of this essay is that the task difficulty has not been discussed. Besides, this is a test deign from my own perspective, which has not been tried in reality. More empirical and reflective test design with a focus on communicative tasks and more pair, group tasks are suggested in the future research.

Funding

This work is supported by "San Quan Education" (Education for all, Education in the whole process, Holistic education) Project "Research on practical approaches of moral education from the perspective of college English teachers-Chinese Culture integrated into moral education" (No.2021xjsqyb05).

References

[1] Rostami. F. and Zafarghandi. A.M., 2014, EAP Needs Analysis in

Iran: The Case of University Students in Chemistry Department, *Journal of Language Teaching and Research*, 5: 924-934.

[2] He, Q**.** 2001, English language education in China, In Baker, S.J., editor, *Language policy: Lessons from global models*. Monterey, CA: Monterey Institute of International Studies, 225-31.

[3] Xiao L X. A new paradigm of teaching English in China: An eclectic model[J]. The Asian EFL Journal, 2009, 11(1): 271-291.

[4] ZHU. H, (2003). Globalization and new ELT challenges in China. English Today, null, pp 36-41.

DOI: 10.1017/ S0266078403004061.

[5] Ur, P. (1996). *A course in language teaching: Practice and theory*. Cambridge: Cambridge University Press.

[6] Luoma, S. 2004, *Assessing Speaking*, Cambridge: CUP.

[7] Lado. R, 1961, *Language Testing*, London: Longman.

[8] Chuang Y Y. Foreign language speaking assessment: Taiwanese college English teachers' scoring performance in the holistic and analytic rating methods[J]. The Asian EFL Journal Quarterly March 2009 Volume 11, Issue, 2009: 150.

[9] Douglas, D. 2010, *Understanding Language testing*, Abingdon: Hodder Education.

[10] Larsen-Freeman, D. 1986, *Techniques and principles in language teaching*, Oxford: Oxford University Press.

[11] Fulcher, G. 2010, *Practical Language Testing*, London: Hodder Education.

[12] Coombe, C; P.Davidson; B. O'Sullivan; & S. Stoynoff. 2012*, The Cambridge Guide to Second Language assessment.* Cambridge: CUP.

[13] Brown, H.D. &P. Abeywickrama, 2010, Language Assessment: *Principles and classroom practices*. New York: Pearson Longman.

[14] Butler F A, Eignor D, Jones S, et al. TOEFL 2000 speaking framework[M]. Princeton, NJ: Educational Testing Service, 2000.

[15] Douglas, D. 2000, *Assessing Language for specific purposes*, Cambridge: Cambridge University Press.

[16] Fulcher, G. 2003, *Testing Second Language Speaking*, Harlow: Longman.

[17] Weir, C. 1993, *Understanding and Developing Language Tests*, Hemel Hempstead: Prentice Hall.

[18] Weir, C. 1990, *Communicative language testing*, London: Prentice Hall.

[19] Bachman, L.F. and Palmer, A.S., 1996, *Language Testing in Practice*, Oxford: Oxford University Press.

Appendix 1

speaking test

1. A sample of the speaking test for senior students of University Y in China

Part 1: one sample of Narration task (this information is provided for both examiners and examinees).

The candidate sees a series of pictures presenting a sequence of events and is required to describe the story based on these pictures in simple past tense (as time suggested in the pictures) for around 2 minutes. One minute is allowed to prepare for the description once the handout is distributed.

(Pictures are absent here)

Part 2: one sample of Structured interview (This information is only for examiners to use, students listen to examiners' instructions in this part)

Warming up-

Question: asking about examinees' personal information (names, nationality, hometown, major, university, hobbies)

Probe-

Question1: Ask the student to explain his/her field to you; what does it involve, what is it about? (Follow-up questions that occur to you.)

Question2: Is the student's chosen filed of study important to his/her country in particular? What is the importance? (if not to her/his country, then to the world in general?)

Question3: What is it about the subject that particularly interested the student?

Question4: How did the student come to be involved in the filed to begin with?

Wind-down:

Question: what is your plan for your future? Where will you go for this weekend?

Appendix 2

Analytic marking scheme (speaking)

This scheme consists of six parts: appropriateness, adequacy of vocabulary for purpose, grammatical accuracy, intelligibility, fluency, relevance and adequacy of content. They are scaled from level 0 to level 4.

Sample of one part is as follows:

Appropriateness

0: Unable to function in the spoken language.

1: Able to operate only in a very limited capacity: responses characterized by sociocultural inappropriateness.

2: signs of developing attempt at response to role, setting, etc., but misunderstandings may occasionally arise through inappropriateness, particularly of sociocultural convention.

3: Almost no errors in the sociocultural conventions of language; errors not significant enough to be likely to cause social misunderstandings.