DEVELOPING THE ASSESSMENT INSTRUMENT OF SPEAKING

A. Ghufran Ferdiant *
Tadris Bahasa Inggris STAIN Pamekasan

Abstract

Speaking means to express ideas orally. By expressing what is in mind, a speaker can make others understand things inside his/her mind. In order to make the others capture and understand what he/she expresses orally, a student should needs to pay attention on the signs that should be fulfilled. How to develop the assessment instrument of the students' speaking ability? Therefore the writer used qualitative research design to describe the way to develop the assessment instrument of the students' ability. The result showed that Developing speaking test is not as easy as other tests because a test developer has to prepare the mechanism or direction and instruction well in order to keep the test valid in which the test developer used content validity to prove that the test was valid. In keeping the reliability the test developer used inter- rater and Pearson Product Moment formula. In fact, content validity, inter-rater and Pearson Product moment formula are proper to assess speaking test. This study will be useful for the English teachers in increasing the ability of the students in speaking by assessing the students' capability in good ways.

Key words: Developing, Assessment Instrument, and Speaking

Introduction

Naturally, students often think that the ability to speak a language is the product of language learning, but speaking is also a crucial part of the language learning process. Effective teachers/lecturer teach students speaking strategies by using minimal responses, recognizing scripts, and using language to talk about language that they can use to help themselves expand their knowledge of the language and their confidence in using it. These teachers/lecturers help students learn to speak so that the students can use speaking to learn.

Language learners who are lack in self-confidence in their ability to participate successfully in oral interaction often listen in silence while others do the talking. One way to encourage such learners to begin to participate is to help them build up a stock of minimal responses that they can use in different types of exchanges. Such responses can be especially useful for beginners.

Minimal responses are predictable, often idiomatic phrases that conversation participants use to indicate understanding, agreement, doubt, and other responses to what another speaker is saying. Having a stock of such responses enables a learner to focus on what the other participant is saying, without having to simultaneously plan a

response. In accordance with the explanation, argues the speaker supplies verbal and nonverbal symbols to the listeners, who receive and interpreted them in terms of their own experiences, beliefs, knowledge, interests, and needs [1].

Some communication situations are associated with a predictable set of spoken exchanges. Greetings, apologies, compliments, invitations, and other functions that are influenced by social and cultural norms often follow patterns or scripts. So do the transactional exchanges involved in activities such as obtaining information and making a purchase. In these scripts, the relationship between a speaker's turn and the one that follows it can often be anticipated. Teachers/lecturers can help students develop speaking ability by making them aware of the scripts for different situations so that they can predict what they will hear and what they will need to say in response. Through interactive activities, instructors can give students practice in managing and varying the language that different scripts contain.

Language learners are often too embarrassed or shy to say anything when they do not understand another speaker or when they realize that a conversation partner has not understood them. Instructors can help students overcome this reticence by assuring them that misunderstanding and the need for clarification can occur in any type of interaction, whatever the participants' language skill levels. Instructors can also give students strategies and phrases to use for clarification and comprehension check.

By encouraging students to use clarification phrases in class when misunderstanding occurs, and by responding positively when they do, instructors can create an authentic practice environment within the classroom itself. As they develop control of various clarification strategies, students will gain confidence in their ability to manage the various communication situations that they may encounter outside the classroom.

Speaking means to express ideas orally. By expressing what is in mind, a speaker can make others understand things inside his/her mind. In order to make the others capture and understand what he/she expresses orally, a student should needs to pay attention on the signs that should be fulfilled. First he/she needs to have an advise, problem, or particular topic in his/her mind in order to convey it to the listeners, neither what should be understood nor responded. Without advise, problem, or particular topic, there will not be a need for him/her to speak. According to Djiwandono, content, organization, and language must get more attention in speaking [2]. If a speaker wants what he/she expresses orally to be able to be understood by other people, he/she has to pay attention on the signs above. The signs are also needed to be criteria for speaking test.

[1] E.E. White, *Basic Public Speaking* (New York: Macmillan Publishing Company, 1984). p. 19.

[2] S.M Djiwandono, *Tes Bahasa* (Jakarta: Indeks, 2008). P. 19.

As with any other area of language assessment, the fundamental issues to be considered in a speaking assessment are: (a) whether or not the test is used as intended, and (b) what its consequences may be (Bachman & Purpura, in press). To ensure that the uses and consequences of a speaking test are fair, the operational definition of speaking ability in the testing context should be examined, since the definition of speaking ability varies with respect to the targeted use and the decisions made. One way to elicit the construct of speaking ability for a certain context is through a scoring rubric which informs test users what a test aims to measure [3]. However, a scoring rubric can affect the speaking assessment, as there may be an interaction effect between the rating criteria and examinees' performance [4]. Different interpretations of the construct may cause biased effects on test-takers' performance, leading to unfairness in scoring and test use. Thus, careful examination of how rating scales interact with speaking performance needs to be considered to determine the fairness of the speaking assessment.

The first issue in examining rating scales is whether the scores given based on the rating scale truly reflect the quality of the test participants speaking performance. Douglas hypothesizes that quantitatively similar scores may not necessarily guarantee qualitatively similar speaking performance [5]. In order to test this hypothesis, the performance of six test participants in a semi-direct speaking test was rated for (a) grammar, (b) vocabulary, (c) fluency, and (d) content and rhetorical organization. The taped responses of test participants were transcribed for qualitative analysis, where the actual language produced by the test participants was described in terms of four rating criteria. Both quantitative and qualitative analyses of test-takers' performance revealed a weak relationship between their quantitative scores based on the ratings and their language production analyzed qualitatively. Meiron and Schick also find that similar quantitative scores represented qualitatively different performance in a role-play simulation task[6]. In their study, the pre- and post-speaking performance of 25 participants in an EFL teacher training program was scored based on a five-category rubric (topic control, pronunciation, grammatical control, lexical control, and conversational control). Close examination of the performance of two test participants, one whose scores increased considerably from pre- to post-test, and the other who exhibited a very small increase,

[3] Sari Louma, *Assesing Speaking* (UK: Cambridge University Press, 2004).

[4] Ibid.; T.F. McNamara, *Measuring Second Language Performance* (London: Longman, 1996).

[5] D Douglas, "Quantity and Quality in Speaking Test Performance (Language Testing)" 11 (1994): 125–44.

[6] B. Meiron and L. Schick. "Rating, Raters and Test Performance: An Exploratory Study" in A.J. Kunnan (Ed.), "Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquiem, Orlando, Florida, Cambridge. UK: Cambridge University Press.

showed that their performances were very different qualitatively, despite similar quantitative scores on their post-test performance. For example, although these two examinees received the same score on conversational control, one examinee's performance showed more of "an academic approach to rhetorical control" while the other's performance exhibited more of "a dialogic approach to conversational control" . The mismatch between examinees' quantitative scores and their qualitative performances, which was found in both of the cited studies, raises questions about the reliability and validity of the testing scores. Thus, for better estimation of test participants' speaking ability, rating scales should be designed to accurately reflect the operational definition of speaking ability[7]. This step can prevent different raters from attending to different features in a test participant's discourse.

What should be considered before deciding on rating scales that ensure the validity of interpretations of test participants' speaking performance? Alderson and Banerjee divides rating scales into two categories. The first category are "generic scales" , which refer to scales that are constructed in advance by proclaimed experts and that are used to evaluate test participants' performance on any type of task. The second category includes rating scales designed to target specific tasks[8]. Rating scales and tasks are thus directly linked because the scales describe the kinds of speaking skills that the tasks elicit[9]. Generic scales have the potential to present inappropriate criteria in measuring the intended ability, a concern related to the issue of validity. Different interpretations of descriptors also lead to problems of reliability[10]. Thus, rating scales developed for particular tasks are more desirable and preferred since they should have greater validity and reliability, particularly those based partially or wholly on a sample of test participants' performance[11].

Another consideration in deciding on rating criteria involves what the speaking test intends to measure. That is, it should be clear what speaking ability means in a given task or test and whether or not defined aspects or features of speaking ability are appropriate for the purposes of the test. Based on criteria used in assessing performance, McNamara distinguished between strong and weak language performance tests. Strong performance tests evaluate test participants' performance based on real-world criteria where how well test-takers perform on a given task is the main

[7] Ibid.
[8] J.C. Alderson and J. Banerjee, "Language Testing and Assessment (Part 2) Language Teaching" 35 (n.d.): 79–113.
[9] Louma, *Assesing Speaking*.
[10] J. Upshur and C.E. Turner, "Constructing Rating Scales for Second Language Tests (English Language Teaching Journal)" 49 (1995): 3–12.
[11] G Fulcher, "Tests of Oral Performance: The Need for Data-Based Criteria" 47 (1987): 287–91; Upshur and Turner, "Constructing Rating Scales for Second Language Tests (English Language Teaching Journal)"; J. Upshur and C.E. Turner, "Systematic Effects in the Rating of Second Language Speaking Ability: Test Method and Learner Discourse" 16 (1999): 82–111.

interest [12] . On the other hand, weak performance tests focus more on the language itself. Such tests attempt to elicit a sample of the test participants' language for evaluation through simulated and artificial tasks, where success of the task is less important than the language elicited.

Although this dichotomy should be understood on a continuum rather than as two separate extremes, McNamara claimed that most general purpose language performance tests are weak in nature [13] . Douglas and Myers questioned what appropriate rating criteria are necessary in a language testing context that has a specific purpose [14] . In their study, they reviewed veterinary students' recorded performances in simulated patient/client interviews. The researchers found out that proficiency was judged according to three different criteria. Participants who were professional veterinarians focused on the test participants' professional relationship with the client and content knowledge, applied linguists concentrated on framework of language use and measurement construct, and student participants used their own knowledge base and the authenticity of the test format. In conclusion, Douglas and Myers argued that raters should blend criteria from different perspectives[15]. Rating criteria derived from task-specific and real-world concerns might not be useful beyond a certain context. Nevertheless, knowledge of indigenous criteria employed in a real-world situation makes it possible to better understand speaking test performance in relation to the situation at hand[16].

Beside that, in education system of Indonesia, the government has stated that there is no grammar minded anymore in studying English, it has been changed into speaking minded because a main success in learning a target language is that the students are able to communicate using the language orally. Nowadays, the syllabus of English in every school all over Indonesia is inclined to focus on how to increase the capability of the students in speaking; however, it doesn't mean that there is no attempt to enhance the capability in other three skills. Hence, it is proper to develop a speaking performance test in which the test participants are the students who are studying English.

To sum up, in order to ensure validity and reliability of a speaking performance test, attention needs to be paid to the quality of the speaking performance along with scoring that is based on criteria specific to that particular testing context. Efforts to ensure high validity and reliability can help guarantee fairness in the speaking assessment. Ultimately, "the point is to get test developers to be clearer about

[12] McNamara, *Measuring Second Language Performance*.
[13] Ibid.
[14] D. Douglas and R. Myers, "Assessing the Communication Skill of Veterinary Students: Whose Criteria? In A.J. Kunnan (Ed.), *Fairness and Validation in Language Assessment: Selected Papers from 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 60-81), UK: Cambridge University Press. 2000.
[15] Ibid.
[16] Ibid.

what they are requiring of test takers and raters, and to think through the consequence of such requirements"

Based on the background knowledge and issues above, the test developer thought that it was very necessary to develop a test in speaking in which it was expected to be beneficial for English lecturers, teachers, students and other test developers.

Research Method

The present study employed a qualitative design to describe the assessment of the students' speaking ability. The subjects were the students of English Department of Teacher Training and Education Faculty, Islamic University of Malang.

Before the test was conducted, the test developer and the lecture assured that the students had known about the objective of the test neither the general objective nor the specific objective. In the specific objective of the test, it had been stated about the aspects that would be evaluated.

The general objective sounds: "The test is to assess the students' speaking skill in expressing ideas orally. " While the specific one sounds:"The test is to assess the students' ability in expressing ideas orally with (1)clear content, (2)well organized, and (3)good language in terms of: intelligible pronunciation, appropriate grammar, appropriately chosen words.

They not only socialized the students about the objectives but also clarified the description of each aspects of speaking competence:

a. Content

The content should be relevant to the topic given in the test. It means that in conveying the spoken text, the whole content of the text should refer to the topic stated by the raters.

b. Organization

The test participant should organize his/her sentences in systematical organization. In other words, he/she should know how to organize the unforgettable experience plot sequences in good arrangement:

Orientation : Tells about whoever were in unforgettable experience plot; what was happening, where and when was it taking place;

Event1 and 2 : Tells about the compilation (either amusing, frightening or embarrassing) and resolution (the way out) in the experience

Reorientation : Tells about the conclusion or ending the event

c. Language

The test participant should be good in the components as follows: grammar, pronunciation, and word choice.

Besides, the test participants were informed that The text told orally would be scored on the following aspects:

Content : 40 %
Organization of ideas : 30 %
Language : 30 %

Total score : 100 %

3.2 Implementation of the Test

On the day of the Test, the raters conveyed the test direction and instruction to keep the test well conducted:

a. Test Directions:

1. All students should be outside of classroom first.
2. They are called one by one randomly
3. Choose one of the topics by lottery.

b. Test Instruction:

Now please tell me about your unforgettable experience when you were ……(related to selected topic) maximum four minutes .

c. Topic and Sub Topics provided

Topic :

Telling about unforgettable experience

Sub Topics:

1. Having a picnic
2. Studying in Senior High School
3. Going camping
4. Watching TV
5. Attending a party
6. Playing a favorite sport
7. Eating a favorite food
8. Helping parents
9. Gardening
10 Making a friend

d. Mechanism of the test:

The test developer and the lecturer were sitting on separated seats while scoring the test participant who was telling his/her experience orally based on the sub topic that had been selected by lottery. In scoring, they scored the test participants based on scoring guide and they determined the minimum difference was 3. Each rater had each own form to write list of score in which the different lists of score from the two raters would be summed and found out the mean.

After the Test

The two sets of the scores from the test developer and the lecturer were summed, then found out the average. They determined the minimum difference was 3. Based on the result of the test there was no score range difference that was higher than the minimum difference they stated. So it was not important for the test developer to use the third rater.

The two sets of the scores from them were summed, then found out the average. They determined the minimum difference was 3. Based on the result of the test there was no score range difference that was higher than the minimum difference they stated. So it was not important for the test developer to use the third rater.

It can be stated that the lecturer had been successful in teaching the students because the average of total score was 8.40 in which the mean score was above the minimum score stated.

Finding and Discussion

Here tells about principles related to selection of test material and test (items) development. There are two principles related to selection test

material and test (items) development based on the opinion of Confucius, they are as follows:

I cannot deny what I experience for myself.

Experience is a part of human destiny. Every human surely has experience neither the interesting one nor the bad one. Moreover, there is an unforgettable experience in which it is very difficult to forget.

Therefore it is not mistaken if the test developer selected this topic as the topic of the test. Telling experience is stated in the syllabus of Speaking 2 course and taught in the Speaking 2 class. Hence, all students as test participants surely knew about their own experience. It can help guarantee the validity of the test

I hear and I know. I see and I believe. I do and I understand.

From the experience a human can hear and know about something. He/she can see and believe. And he/she can understand something by doing. In short, every human can get lessons from the experience.

"I do and I know" refers to the competence of human. So here the test developer assessed the students or the test participants' speaking skill, especially in telling about unforgettable experience. Because there are various stories in experience, thus he provided ten sub topics related to experience to be selected. Besides, he intended to avoid that the test participant would inform one another about the test.

Validating the Test

Language test can be defined as a means or procedure used to evaluate learning process. The test should refer to measure the language ability possessed by the test-taker or the test participant. Related to language test, Djiwandono states that in language learning implication, a test is intended to measure language competence as the reflection of learning result. In addition, he states that a good test should have some characteristics, two of them are validity and reliability[17].

Validity

To prove the validity of the test, the test developer used ***curricular validity*** in which the validity could be proven from relevancy between the test and the curriculum used in the department.

To keep ***validity*** of the test, before scoring the test developer asked the lecturer about the materials having been given to the students. The test developer and the lecturer agreed that the topic of the test was unforgettable experience in which in telling the unforgettable experience a student had used one of the sub topics selected by lottery. Moreover, the test developer provided Table of Specification in order to guard the relevance between the test and the

[17] Djiwandono, *Tes Bahasa*. P. 163.

objectives of the test neither the general objective nor the specific objective.

After scoring, it was found that the speaking test given to the students was relevant to of the test neither the general objective nor the specific objective. It means that the test given to what the lecturer had explained in the speaking 2 class. Beside that it was relevant to the objectives. It means that the test given to the students was valid.

Reliability

While to keep ***reliability*** of the test, before scoring the test developer and the lecturer agreed to implicate Inter-Rater Reliability in which in considering the reliability level there should be two lists of score toward the test participants obtained from two raters. It was stated that the test developer was as the first rater and the lecturer became the second rater. Besides, He gave the lecturer scoring guide in order to make the process of scoring more reliable (there is consistency in scoring).. In other words, It was expected that there was no distinction in scoring the student's ability in speaking, especially in telling unforgettable experience. In addition it was agreed that the minimum score or passing score was (2+2+2 = 6) and the minimum difference was 3.

After scoring, there were two lists of score toward the test participants obtained from two raters in which they scored based on the scoring guide provided. It was stated that the test developer was as the first rater and the lecturer became the second rater. Based on the two list of score, It can be assured that there was consistency in scoring the ability of the students to tell unforgettable experience orally. In other words, the test was reliable. It could be seen from the difference between the first rater and the second rater in scoring. The minimum difference stated was 3. Whereas, the highest difference was only 2.

Moreover in proving the reliability, the test developer used the formula of Pearson Product moment. The two different lists of score was processed using the formula in order to know the reliability of the test.

Conclusion and Suggestion

Developing speaking test is not as easy as other tests because a test developer has to prepare the mechanism or direction and instruction well in order to keep the test valid in which the test developer used content validity to prove that the test was valid. In keeping the reliability the test developer used inter-rater and Pearson Product Moment formula. In fact, content validity, inter-rater and Pearson Product moment formula are proper to assess speaking test.

In developing speaking test it is better to employ content validity, inter-rater and Pearson Product moment formula because they can work well in keeping the test valid and reliable.

References

Alderson, J. C., & Banerjee, J. 2002. Language testing and assessment (Part 2). *Language Teaching*, *35*, 79-113.

Bachman, L. F., & Purpura, J. E. (in press). Language assessments: Gate-keepers or door openers? In B. M. Spolsky & F. M. Hult (Eds.), *Blackwell handbook of educational linguistics*. Oxford, UK: Blackwell Publishing.

Djiwandono, S.M. 2008. *Tes Bahasa*. Jakarta: Indeks

Douglas, D. 1994. Quantity and quality in speaking test performance. *Language Testing, 11*, 125-44.

Douglas, D., & Myers, R. 2000. Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 60-81). Cambridge, UK: Cambridge University Press.

Fulcher, G. 1987. Tests of oral performance: The need for data-based criteria. *English Language Teaching Journal, 41*, 287-91.

Jack C. Richard. 2000. *Interchange*. UK: Cambridge., authentic materials

Luoma, S. 2004. *Assessing speaking*. Cambridge, UK: Cambridge University Press.

McNamara, T. F. 1996. *Measuring second language performance*. London: Longman.

Meiron, B., & Schick, L. 2000. Ratings, raters and test performance: An exploratory study. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* . Cambridge, UK: Cambridge University Press.

Renshaw, J.2008. *Boost Speaking 2. Hongkong*: Pearson Longman Asia ELT

Upshur, J., & Turner, C. E. 1995. Constructing rating scales for second language tests. *English Language Teaching Journal, 49*, 3-12.

Upshur, J., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*, 82-111.

White, E.E. 1984. *Basic Public Speaking*. New York: Mcmillan Publishing Company.