タイトル	Some issues, options and recommendations in the testing of spoken interaction for students of oral and general English at universities in Japan
著者	Munby, Ian
引用	北海学園大学人文論集, 28: 133-145
発行日	2004-07-31

Some issues, options and recommendations in the testing of spoken interaction for students of oral and general English at universities in Japan

Ian Munby

Abstract

Testing spoken interaction effectively in the language classroom presents the teacher with a number of complex challenges. These challenges may be further complicated in university classes in Japan when testing large numbers of students of similar level. This paper aims first to identify some important issues involved and options available in such tests including test design, assessment, and implementation, using data from questionnaires completed by 25 teachers of oral or general English in colleges in Hokkaido regarding their preferred methods of testing spoken language performance. In support, I draw on my own experience testing college-level students and as an oral examiner for public English examinations and also from the literature on language testing theory. Finally, I include some recommendations for practitioners and an example test.

日本の大学におけるスピーキングテストについての問題点・選択肢・提案

英語の授業でスピーキングテストを効果的に行うことは教師にとって非常に難しい。日本の大学において，大勢の同等レベルの学生の場合ではさらに困難である。この論文の目的はそれに関係する重要な問題点を明らかにし，いくつかのテストの例を提示した。内容は，北海道の大学の英語教師25名による彼らが好んで用いるスピーキングテストに関してのアンケートのデータを用い，テスト様式，実行方法，評価方法を収めた。それをサポートとするものとして，大学生にテストを行った自分自身の経験，公的な英語試験のスピーキング試験官としての経験，言語テスト理論についての文献を利用した。最後にテストについてのアイディア・提案を示した。

What is a speaking test?

By definition, a speaking test is a test which requires the candidate to engage in spoken interaction with another candidate, candidates, or examiner, to complete communication tasks with pre-specified goals within a time limit, allowing the examiner to assess and grade performance. For the purposes of this paper, a speaking test does not refer to "speech-making" tests or oral presentations.

Issue One. To test or not to test?

Ten respondents (N=25) to the questionnaire (see Appendix 1) preferred not to give a speaking test, with four key reasons emerging. Four respondents mentioned that classes with large numbers of students (50-60, for example) rendered the implementation of such tests practically impossible. Some mentioned that, even in smaller classes, reliable grading was unachievable. Others claimed that the level of their students' speaking ability and motivation was too low. Finally, two respondents believed that the pressure and artificiality of the testing environment resulted in their students being too nervous to perform to the best of their abilities.

Among reasons for giving a speaking test cited by the remaining fifteen "test-doers", the most common was that if we teach oral communication, we should test it. However, three respondents pointed out that accurate assessment was not the main goal of the speaking test and that pushing the students, motivating them, or giving them a reason to make an effort to improve their speaking skills in course work was more important.

Option

However, three respondents stated that they preferred to regularly "listen in" on students as they engaged in oral task work in regular class time and quietly award grades in what was described as a "cumulative" process of evaluation. This could be done in addition to, or instead of, a formal speaking test.

Issue Two. How do we design a test?

Unfortunately, very few general English or speaking skills course books provide ready-made speaking tests which are suitable for testing our students and teachers are left to design their own. In the questionnaires, most test-doers described their tests as consisting of between two and four distinct stages with combinations of the following tasks: conversation, question and answer, role play, information gap activities, task-based dialogues, topic-based discussion, and telling a story from pictures.

As a general rule of thumb, speaking test tasks should not only reflect tasks covered in the course, but should also reflect either tasks encountered in real life, or have some "lifelike" face-validity, or relevance to the students' lives.

In order to construct a test of spoken language, Weir (1993) suggests first considering available theories on what is involved in the speaking skill and, second, defining the operations, performance conditions, and expected quality of output. Regarding operations, Bygate, cited in Weir, suggests that speakers draw on a repertoire of routines in order to complete a communicative task and a logical point of departure in test construction would therefore be the identification of the routines enacted to complete the type of task being tested. Weir notes that operations involve both informational and interactional routines and improvisational skills.

For an example test written by myself, see Appendix 2. In Part One of this test, interactional routines involve asking questions and giving responses relevant to the making of arrangements, and for finding out personal information and opinions in Part Two. Informational routines concern ways participants present information relevant to task completion. In the event of breakdown in the interaction, what Weir, drawing on Bygate (1987), describes as "improvisational skills" are enacted and involve "negotiation of meaning" and "management of interaction".

According to Weir, conditions affecting performance on the test include processing under normal time constraints, purpose, and nature of interlocutors.

Processing under normal time constraints. For the example test, the time limit of four minutes for each part was based on the performance of moderately proficient test-takers. Data from the questionnaires indicate that while one test-doer employed a 1:4 format in a fifteen minute session, most adopted a ten-minute testing session with either pairs or groups of 3 or 4. The number of students each test-doer claimed to be able to test in one ninety-minute session ranged from 8-40, with an average of 21.

Purpose. The purpose of the example test is to measure the student's:

1. course content mastery.
2. ability to interact effectively in English to complete a task, or negotiate a task outcome.
3. speaking skills and level of communicative competence in English

Nature of interlocutors. Weir suggests that the number of participants in the interaction, their status, and familiarity may affect performance. First, regarding preferred format, among the fifteen test-doers, only two reported adopting a 1:1 format, with some mentioning having abandoned the practice since it constrained the examiners ability to rate effectively, took too much time, did not reflect the way students had practiced oral communication in class, and made the students nervous.

On the other hand, with a 1:1 format, a more level playing field can be established if the examiner's input remains more or less standard. Furthermore, the examiner is able to operationalize the testing of improvisational skills, through feigning misunderstanding or asking for clarification. With the 1:2 or 1:3 situation, opportunities for candidates to demonstrate these skills simply may not occur. A further disadvantage of the two- or three-peer format is that a weaker candidate may affect the performance of a candidate with superior language skills, or vice versa, a significant performance condition variable noted by Hughes (1989). Regarding status and familiarity, the candidates would normally know each other and share equal status as members of the same class.

Recommendation

Seven respondents claimed to have received training in, or worked as examiners, in the one or more the following public English examinations: Eikken, IELTS (International English language Testing System), FCE (Cambridge First Certificate of English), and SST (Standard Speaking Test). In addition, some teachers appeared to be strongly influenced by these tests, or even adopted them wholesale for college classroom use. It is certainly advisable for practitioners to familiarize themselves with as many of these tests as possible, and, preferably, take rater training courses to hone their assessment skills.

Issue Three. How do we assess performance?

Weir describes quality of output as "the expected level of performance in terms of various relevant criteria" (1993. p30) and these are usually represented as assessment criteria, or "band descriptors for grading". Fluency, or smoothness of execution, was the most commonly mentioned element of spoken language performance assessed by the 15 test-doers. Other criteria mentioned were: grammatical accuracy or structure, task achievement, discourse management, information given, discussion skills, communication strategies, participation, quality of interaction, attitude, vocabulary, range of language, and pronunciation. Most test-doers appear to either rate up to four of these on individual sub-scales, or mentally combine them into a single unitary scale. Three test-doers admitted having given up using sub-scales since they were destined for eventual combination anyway, but the final distribution of grades on the scales they employed ranged from A, B, C, with D as fail, or from 4-10, with 4 as fail, my personal preference. However, 12 test-doers claimed never to fail students on their test performance, with three mentioning that failure would only result in combination with other factors such as poor attendance or poor performance in other areas. For the majority of test-doers, speaking test grades made up 20-30% of the course grade.

What follows is the bad news regarding assessment. First, some commentators such as Matthews (1990) have dismissed attempts to measure proficiency by reference to behavioral criteria as basically flawed, preferring non-linguistic task achievement, such as the degree of success with which candidates can ask for street directions and mark a place on a map. Second, assessment of spoken interaction cannot be said to be reliable unless performances ratings are checked for intra-rater reliability, or the ability of the examiner to re-rate a recording of a performance and assign the same score. When testing in a team, inter-rater reliability, the extent to which raters agree on scores, is also essential. In this case, sub-scales are useful for identifying sources of variation.

In view of Fulcher's (1987) sobering comment that if the test is not proved to be reliable, we are not actually measuring anything at all, what constitutes reliable assessment? The "rock-bottom" reliability co-efficient is quoted at STD, or standard deviation, 0.8 by Upshur and Turner (1995) and at 0.7 by McNamara (2000). My only experience of re-rating college speaking tests was with a doctor-patient role play test for the Medical English Course at Hokkaido University (Holst. 2000). Regarding intra-rater reliability, using a scale of 4–10, my re-rating of 32 candidates produced a co-efficient of 0.76, while with inter-rater reliability the correlation was 0.59. It is worth bearing in mind that in the IELTS test, for example, gaining one point on a scale of 0–9 from non-speaker to native-like speaker requires between three months and six months of full-time study. Clearly, measuring our students' speaking abilities reliably in a class of students of similar level and learning experience on a scale of 4–10 is very difficult. While reducing the number of levels on the scale to A-D might appear to be a solution, these letter grades will probably have to be transformed into pure numerical scores anyway if they represent a percentage portion of the final grade.

Recommendations

A further problem concerns college freshmen beginning two-semester courses of oral language instruction with failing levels of fluency. Since the 35-45 hours of available class time may not be enough to rectify the situation, I find that teaching and measuring speaking strategies, such as strategies for initiating conversations, or keeping a conversation going by asking follow-up questions, can be more effective. Furthermore, raising our students levels of socio-pragmatic awareness can also lead to measurable gains in the communicative competence of our students. With reference to Part One of the example test, I pay attention to, and rate, the pragmatic quality of interactive operations. For example, in making arrangements for a trip to the movies, "I say we see *The Last Samurai*" is rated low as an initial suggestion, compared with the more appropriate "How about *The Last Samurai*?".

Regarding grading, in the testing room, I try to award two numerical scores, 6 or 7 for example, to my students either before or very early on in the test. The process of evaluation is therefore a process of choosing which of the two scores to award. In the event of being unable decide, I leave it at 6/7, or 65%.

Issue Four. How do we implement the test?

Almost all test-doers give the speaking test at the end of the course, with some preferring to test mid-course as well. Most organize the test by assigning time slots for groups or pairs and let the students know what they will expected to do before the testing day.

Options

There are a number of options here. A few test-doers choose the pairings or groupings themselves, but the majority either pull names out of the hat on the testing day or invite the students to choose their own partners in advance. The latter allows for optimal performance conditions by giving the students the freedom to interact with people they know, like and trust. However, even when the test tasks are decided by the roll of the dice, as in the example test, I have noticed that this arrangement allows the students to predetermine task outcomes, thereby defeating purpose 2 of the test mentioned earlier.

Recommendation

First, make sure that the students choose different partners in every lesson so that they are well prepared to be tested with anyone. Second, make a note of who you think would make good testing partners by observing social behavior during the course, but only announce the pairings or groupings moments before inviting them into the testing room. If possible, allocate two 90-minute sessions for testing a group of 25-30 to allow some time for personalized conversation at the beginning of the test, before rolling the dice, and at the end. I feel this to be especially important in view of the fact that my students most commonly reported English course goal is "to be able to talk with foreigners". If you, the examiner, are just sitting their silently rating the students, they may well be disappointed. Nevertheless, use a kitchen timer to ensure that the total time allocated to each test remains constant.

Conclusion

Despite all the complications, especially regarding rater reliability, I would recommend teachers to go ahead and test without fear, remembering that the college speaking test is not a high-stakes career-determining test and can indeed be fun.

Appendix 1. Questionnaire

1. Do you test your students' speaking skills in a formal test of spoken interaction?
2. Do you test their English speaking ability in some other way, such as presentations?
3. What are your reasons for giving or for not giving, a speaking test?

If you do give a speaking test…..

4. When do you give the test?
5. What percentage of the final grade does the test result represent?
6. Do you ever give students failing grades for poor performance in the speaking test?
7. What format do you use? Why?
8. How much time do you allow for each test?
9. How many students can you test in one ninety-minute session?
10. What kind of tasks you set for your students?
11. Do the students know what kind of tasks they will do before the test-taking day?
12. Please explain your test-taking procedure.
13. What are your assessment criteria?
14. Have you changed the way you do speaking tests over the years?
15. Can you provide example materials or scoring criteria?
16. Finally, have you ever done any oral testing work for public

examinations?

Have you done any oral examiner training courses?

Appendix 2. Example speaking test for college freshmen.

Test Procedure

1. There will be two parts to the speaking test. Part I. Situations, Part 2. Discussion
2. Your task will be decided by the roll of a dice.
3. Each part lasts four minutes.
4. You may NOT use notes.
5. Your teacher will only listen to you.

Situations (Role Play) You are friends.

Make a plan for the following.

1. A shopping trip
2. A camping trip
3. A weekend trip to Tokyo
4. A one week trip to a foreign country
5. A trip to the movies
6. A birthday party

Discussions

1. Find out what your partners did at the weekend.
2. Find out what your partners are planning to do in the summer holidays.
3. Find out what your partners like doing in their free time.
4. Find out about some interesting places your partners have travelled to.
5. Find out about your partners' favorite restaurants.

6. Find out which city your partner likes best: Otaru or Sapporo.

Grading

Each student will be assessed individually. There is no group score.

Grade A (80-100%) Participates actively and effectively in the speaking tasks. Very little hesitation and pausing. Easy to understand. Extended utterances. Interacts well with others. Uses grammar and vocabulary effectively.

Grade B (60-80%) Generally participates effectively in the speaking tasks. Some hesitation and pausing. Generally easy to understand. Some extended utterances. Generally uses grammar and vocabulary effectively.

Grade C (50-60%) Sometimes participates effectively in the speaking tasks. Frequent hesitation and pausing. Sometimes difficult to understand. Short, simple utterances. Sometimes uses grammar and vocabulary effectively.

Grade D (40-50%) Fail. Hardly participates in the speaking tasks. Frequent breakdowns in communication. Very difficult to understand. Uses very few words in each utterance. Does not interact with others. Does not use grammar and vocabulary effectively.

References

Fulcher, G. 1987. "Tests of oral performance; the need for data-based criteria." ELTJ. 41 (4): 287-291.

Holst, M and R. Evans. 2000. Medical English. An ESP Application.

JALT Hokkaido 2000 Proceedings. JALT Hokkaido. Sapporo.

Hughes, A. 1989. *Testing for Language Teachers.* Cambridge Handbooks for Language Teachers. Cambridge. Cambridge.

Matthews, M. 1990. The measurements of productivity skills: doubts concerning the assessment criteria of certain public examinations. ELTJ 44 (2) 117-121.

McNamara, T. F. 2000. *Language Testing.* Oxford Introductions to Language Study Series. Oxford. Oxford.

Upshur, J and Turner C. 1995. Constructing rating scales for second language tests. ELTJ 49 (1): 3-12.

Weir, C. J 1993. *Understanding and Developing Language Tests.* London: Prentice Hall International.