Creating an Advanced Level Speaking Test

By
Sarah FAHERTY*

Abstract

This extension to the standard speaking test was designed to assess the speaking level of higher tier students in second year university classes, using the CEFR-J as a basis. It was designed to complement the existing speaking test and to enable students to demonstrate skills up to B2 level of the CEFR-J. The test was designed to fit in with the existing curriculum using topics covered within it. The validity of the test was established with reference to Bachman and Palmer's 1996 model for establishing test usefulness.

Key Words:　Speaking, CEFR-J, Assessment

1. Introduction

This extension to an existing speaking test was developed at a small science and engineering university in west Japan. All students at this university take communicative English courses in the first and second years. The final assessment in two of the four semesters of the English communication curriculum is a speaking test. In this test, two students discuss a topic for three minutes, with ten seconds to think before beginning. These speaking tests are designed to assess general communication skills in conversation about topics such as hobbies and daily life, which are covered in classes. They test up to B1 on the CEFR-J. While these tests are a relevant assessment of performance in English Communication, a need to further assess students beyond B1 level was recognized. The intention was not to replace the standardized test, but instead, to complement it, and to offer students performing at B1 level and above an opportunity to demonstrate their skills by assessing their discussion skills on a range of topics that are both related to the curriculum and relevant to the lives of university students. In 2016, three groups from the Pharmacy and Architecture and Mechanical Engineering departments, completed the test, examined by one of the test developers.

Theoretical Background

There are many things to consider when developing a test, in order to ensure its validity. As

* Senior Assistant Professor, Sojo International Learning Center

Bachman and Palmer (1996) discuss, misconceptions about language testing can lead to inappropriate tests for the situation in which they are being used. Having unreasonable expectations of what a language test can do, or placing blind faith in the test can lead to the use of assessments that are inappropriate for the test takers (Bachman and Palmer, 2996: 3). To avoid these potential pitfalls, Bachman and Palmer (1996) recommend a model of test usefulness to act as quality control. Their definition of test usefulness includes six items: reliability, construct validity, authenticity, interactiveness, impact and practicality.

Test Usefulness

Reliability

Reliability describes the consistency of the results (Luoma, 2004). If the same test was conducted with same people on another day, the results should remain largely the same. This is particularly important in this case, because multiple assessors are conducting the test. For large scale tests conducted by exam boards, lengthy selection and training programs are conducted.

This was deemed unnecessary in this context, as the assessors were also the test developers. Additionally, an adapted version of the existing rubric, which assessors were already familiar with, was used.

Construct Validity

Luoma (2004) argues that validity is the most important quality in test development, as it refers to the meaning of the scores. Bachman and Palmer (1996: 21) state that it "pertains to the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores". It is important to demonstrate that the test scores produced by the test actually reflect the area of language ability we want to measure.

Authenticity and Interactiveness

Bachman and Palmer (1996) argue that these are not separate properties, but are part of construct validity. They argue that it is an important quality of the test because the language used in the test should correspond to the target language situation the test is intended to replicate. Authenticity will also affect participants' perception of the test (Bachman and Palmer, 1996). Test takers are more likely to perform at their best in the assessment if they can perceive the relevance of the language being assessed. This is related to interactiveness, which Bachman and Palmer define as "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (Bachman and Palmer, 1996: 25). These characteristics include the learner's language ability, topical knowledge and affective schemata. It is possible for a task to be highly authentic, by accurately recreating a task likely to be used in the participants' lives, but have low interactiveness, by not requiring the participants to process the language, and merely repeat it. In order to determine the interactiveness and authenticity of a test task, three sets of characteristics must be considered, the characteristics of the test takers, the characteristics of the target language use and the characteristics of the test task.

Impact

The element of impact that was most important for this test in particular is washback (Bachman and Palmer, 1996). Washback is the effect of assessment on the course, and can be either positive or negative. If teachers feel obliged to teach to the test, to the exclusion of other language skills, that are not being assessed, the washback is negative. However, if the purpose of the test is to broaden the curriculum, and encourage further development of language skills, the test has a positive impact, or washback.

Practicality

For a test to be useful, it must be practical. The resources needed to administer the test must be available (Bachman and Palmer, 1996) otherwise the test cannot be implemented effectively. A balance between practicality and the other test characteristics described above is essential.

Test Development

Test Format

The current standardized test comprises a two person conversation lasting two minutes and forty-five seconds. Changing the format for this extension presented an opportunity for students to develop further communication skills, so, after consulting the CEFR-J, a discussion based test was designed. The CEFR-J can-do statements for spoken interaction at B1 level states that the learner can agree and disagree politely, exchange personal opinions, negotiate decisions and ideas (Nagai and O'Dwyer, 2010). A learner operating at B2 level of spoken interaction should be able to participate in an extended conversation about a subject of personal, academic or professional interest and be able to explain his or her opinion with relevant arguments or comments. As a result, topics such as "What are the pros and cons of living alone?" or "What are the pros and cons of having a part time job?" were used. It was anticipated that the use of such topics, of relevance to Japanese university students and based on topics in the curriculum, would increase the positive perception of the test by the learners, as well as leading to positive washback in the form of further discussion based activities in the class.

To provide further opportunity for expansion, different groupings were considered. Although there are a number of group projects within the curriculum, there are few opportunities for students to develop the skills needed to discuss ideas in groups, rather than pairs. As a result, groupings of three people were chosen. It was hoped that this would improve student performance in group projects. Additionally, having larger groups allows for a longer test and also provides an opportunity to truly test discussion skills. In trials, it was observed that participants each made a short speech, rather than asking and answering questions.

In order to encourage longer discussion, the test lasts four minutes, with ten seconds to think about the topic before beginning. The longer test leaves enough time for learners to discuss the opinions they have stated by asking and answering questions in the intended manner. It is also anticipated that a washback effect will lead to further development of the second year curriculum in particular, with the addition of discussion based activities, and activities that vary beyond the usual pair groupings.

Rubric

An adapted version of the existing rubric was used, as testers were already familiar with it. This rubric consists of three main areas for assessment: fluency, lexico-grammar and interaction skills. The requirements for fluency and lexico-grammar were found to be similar between this test extension and the standard test. As a result, these sections were used as they were. Some small changes were made to the interaction skills section, as this was found to be different between the two tests. Further minor edits were made to the interaction skills section following the test to improve clarity. As a result of using this established rubric, little training was needed prior to the test. This will be advantageous should use of the test expand to other classes as anticipated, because all teachers in the department are familiar with it.

Norming

As the developers of the test were the examiners,

a formal norming process was deemed unnecessary. However, video recordings of two groups were made, and could be used for future norming to ensure content validity.

Student feedback

In total, 69 students completed a survey about the test. Overall, the response was positive. 94.2% of the students surveyed agreed with the statement, “The discussion test is an important part of EC4” and 92.75% believed the test had helped them to improve their English skills. Furthermore, 91.3% agreed that the test was a good way to test their English skills. This positive feedback reinforces the benefits of this test, and suggests that it should be continued in future years.

In terms of the test format, 63.77% of the students surveyed disagreed with the statement, “The test was too long”, and 36.24% agreed. Although the majority of students disagreed with the statement, the reasons for agreeing should be considered. Additionally, 39.13% of students did not express satisfaction with their performance in the test, compared to 60.87% who did. It should be considered that the students who felt the test was too long may have felt so because they felt ill-prepared for it. Further research needs to be conducted on this point. 75.36% of the respondents stated that the topics discussed in the test were interesting for them. However, 24.64% were not interested in the test topics, which, while a minority of students, is still a fairly significant number. Before the next test, alternative topics could be piloted and the feedback compared.

The final comment section of the survey generated little feedback, however, one commenter suggested providing feedback to students following the test. As the test is graded instantaneously, it should be possible to provide this feedback in the future, although a suitable format needs to be considered.

Further research

As has already been stated, it is hoped that this test will continue to be used in the future, with a larger number of students in different departments. This may require greater training for markers, as well as a norming process to ensure consistency. A video of varying proficiencies of student that could be used for this purpose was taken of two classes. Additionally, it is expected that involving other teachers would generate greater discussion about the test, and may lead to further editing of the rubric and test format.

Additionally, further preparation for the test within the current curriculum would benefit students greatly. In the academic year 2016/17, a group of teachers from the department will be building a bank of speaking tasks to fit in with the current curriculum. As part of this project, discussion based activities designed to give students experience with the format of the test, as well as appropriate language and strategies, will be developed. It is hoped that providing experience of the test format will lead to increased satisfaction with their own test performance, not least because students will have a clearer idea of what is expected of them.

These speaking tasks will also allow teachers to assess the suitability of this test for their classes. In the 2014/15 academic year, many teachers felt that the test would be too challenging for their classes, and would prove to be demotivating to otherwise motivated students. While the extension requires a new skill for most students, it is not necessarily linguistically challenging for higher level learners. By using the speaking tasks with their learners, teachers may have more confidence in opting to include the test extension as part of the assessment for their classes.

Finally, methods of feedback will be considered, following a comment from a student that they would have liked more information about their performance and areas for improvement. Although

the data is collected soon after the test has been completed, an appropriate way to disseminate this data needs consideration, and should be accompanied by useful advice about working on these areas for improvement in order to make the feedback as beneficial as possible.

Conclusion

While there are undoubtedly improvements to be made with this test, which have already been discussed, it has the potential to offer a useful challenge for our higher level learners. As such, it is of benefit to the second year students in higher tier classes. Moving forward, expanded use of the test, with students from a variety of faculties and levels, will assist in further improving the current format and verifying its reliability.

Acknowledgements

The author would like to thank Hana Craig and Jon Rowberry for their contributions to this project.

Bibliography

Bachman, L. F. and Palmer, A. S. (1996). Language Testing in Practice. Oxford: Oxford University Press.

Luoma, S. (2004). Assessing Speaking. Cambridge: Cambridge University Press.

Nagai, N. and O'Dwyer, F. *CEFR and ELP*. Retrieved from FLP SIG Framework and Language Portfolio SIG Establsihed Within JALT: https://sites.google.com/site/flpsig/flp-sig-home/language-portfolio-for-japanese-university

O'Sullivan, B. (2011). *Language Testing: Theories and Practices*. Basingstoke: Palgrave Macmillan.