

大学入試における英語スピーキングテストの基本的問題 Theoretical and Practical Issues of Assessing Speaking in University Entrance Examinations

中村 優治 NAKAMURA, Yuji

● 慶應義塾大学
Keio University

1. Purpose of the Paper

While an integrated approach to language and content learning and instruction has recently been promoted, the aspect of integrated language and content assessment has not been sufficiently discussed. The purpose of this paper is to discuss fundamental issues when implementing a speaking section in a university entrance examination in terms of an integrated test. These issues will be raised and discussed from the following four viewpoints: a) benefits of introducing the speaking assessment section in the university entrance examination; b) deficiencies of not having the speaking assessment section in the entrance examination; c) theoretical problems for constructing the speaking test; d) practical problems for implementing the speaking test

2. Testing objectives and learning activities

Brown (2012) claims that an important role of teachers in any placement, diagnostic, progress, or achievement assessments should be to insure that these assessments match the things students are

learning and practicing in their classroom activities. The connection between assessment and instruction should be taken into account when making an assessment. In other words, what are students learning? How well are they learning it? How can instruction support learning? All of these factors should be considered in assessment. Accordingly, the result of linking assessments to classroom activities will much more closely match those assessments to the belief systems of the teachers, the syllabi being used to organize the teaching, and the techniques and exercises that teachers are using to foster learning in their classrooms (Brown, 2012).

3. Teaching and learning objectives

The objectives of the Course of Study for Japanese high school students proposed by the Ministry of Education, Culture, Science and Technology are “Based on general instruction to develop basic communication abilities given in lower secondary schools, the four areas of language activities should be integrated for comprehensive learning, while incorporating appropriate language activities involving speaking and writing based on what students have heard or read” (MEXT). The gist of

this objective is as follows:

- a. Understanding information, ideas, etc., and grasping the outline and main points by listening to introductions to specified topics, dialogues, etc.
- b. Understanding information, ideas, etc., and grasping the outline and main points by reading explanations, stories, etc. Reading passages aloud so that the meaning of the content is expressed.
- c. Discussing and exchanging opinions on information, ideas, etc., based on what one has heard, read, learned and experienced.

In the MEXT statement, the integration of the four skills is highlighted strongly. This is aligned with the concept of the *integrated assessment* in which four language skills ought to be integrated in the assessment. The integrated assessment will be discussed in detail later.

Because communication does not take place using only one language skill, the separation of the four skills in instruction seems less effective (Henkel, 2010). For learners to attain language competence, it seems that teaching must integrate linguistic and communicative skills. As Henkel (2010) claims, as the goal of instruction is to advance learners' language proficiency required for communicating effectively, integrated language teaching will likely continue to dominate the types of pedagogical models available. These are completely aligned with the MEXT statement mentioned above.

4. Test development and an integrated assessment

4.1 A new test development

When developing a new test, especially performance tests (i.e. speaking and writing test), special attention should be paid to the following

ideas in relation to language tests:

- a. What is to be assessed (construct and content)
- b. The employment of a range of tasks to provide test takers with opportunities to perform their best (test method)
- c. Indication of how the performance is scored (rating criteria and raters)
- d. Explanation of how the test result is interpreted (score interpretation)
- e. Separate scores must be reported for each task, not a single combined score (test result reporting)
- f. The development of locally sensitive instruments (practicality) (cf. Council of Europe, 2001; O'Sullivan and Nakatsuhara, 2011)

4.2 Integrated assessment

By taking the above items into consideration, we will look into Plakans (2011, 2012 and 2013) ideas which can provide detailed description and thorough discussion on integrated assessment. Plakans (2011) discusses integrated assessment as follows:

First, she defines integrated assessment as the use of tests that combine two or more skills, such as reading/writing or reading/listening/speaking. Usually the receptive sections lead to a final (written or spoken) performance section. There are various relationships between the receptive and the productive sections: the performance could be a summary of or an opinion on the topic, or it may simply involve responding to a related topic, whereas the receptive sections are used for inspiration.

Integrated tasks require examinees to integrate multiple language skills in a substantial way to complete whatever speaking task is at hand, e.g. to understand academic texts and lectures and create spoken responses that demonstrate understanding of those texts and lectures. "Integrated skills" tasks thus require test takers to use the information

provided in the reading and listening passages in written and/or spoken responses.

Second, Plakans (2011) discusses the difference between integrated and individual skill testing. She claims that integrated tests are more complex, containing tasks requiring “multiple steps or a longer process.” Independent speaking tasks are tasks based on a stand-alone prompt or visual, while integrated tasks involve both listening and speaking, or both reading and speaking. While the integrated tasks provide the information about which examinees will speak, the independent tasks usually require examinees to rely on their personal experience or general knowledge to complete the task.

Integrated tasks are more complicated to construct, requiring that test-makers first develop an input task and then ensure that the performance task is appropriate for the input. Plakans also states that the scoring may be different; it could rely at least in part on an assessment of whether the test-taker has used the input material appropriately.

Third, Plakans (2011) introduces several methods of integrated assessment. She recommends the use of multiple measures, i.e. pairing integrated tasks with independent tasks, or utilizing more than one integrated task. She also emphasizes the importance of pilot testing, as both test takers and raters are likely unfamiliar with this type of test format. It is important to look at how integrated assessment is handled.

Fourth, Plakans (2011) describes the challenges of integrated assessment. She indicates that the main issues are: 1) the construct of speaking ability, 2) the tasks (test tasks and response tasks), and 3) the scoring and raters. It is important that the skill an integrated task measures be considered. It is crucial that, as the test format is unfamiliar, instructions be clear and succinct. When rating examinee responses from integrated tasks, raters must attend to content accuracy in order to ensure that the examinees have adequately understood what is presented in the test

or lecture. Additionally, raters must consider the source material when scoring.

Finally, Plakans (2011) stresses the benefits of integrated assessment. Plakans claims that, in spite of the complexity and difficulty in its nature, integrated assessment has the following benefits: 1) Integrated assessment is less reliant on the ‘four skills’ model of language and appreciates that language is not a unitary construct; language skills interact with one another; 2) Integrated assessment has more authenticity: these tasks are similar to those encountered in real-world language use; 3) With an input task as inspiration, performance tasks in integrated assessment are easier than in independent tasks; Test takers show more confidence in their performance having read or heard something related to the topic; 4) Students receive positive wash-back effects from these test tasks.

Obviously, Plakans (2012) has included thoroughly all the necessary concepts for test construction, such as test validity, reliability, practicality, authenticity and impact.

As Plakans (2012) states, language, when viewed as a whole concept comprised of the four skills, allows performance to be viewed more substantially and in terms of real world ability. However, those tests are more difficult to construct than their counterparts which test isolated abilities. Speaking tasks that require listening to and then giving a speech evaluated on explicit criteria is one such example of an integrated test task. With foundations in integrated assessment, integrated skills assessment is not a new idea. However, development and use of such tests is challenging and research is ongoing (Plakans, 2012).

5. Technological assistance in a new test

In addition, two other aspects should be strongly stressed in this day and age of information technology. Davidson and Coombe (2012) state

that computerized testing and web-based testing cannot be ignored. Computerized testing can help us produce more valid and reliable tests that are more accurate measures of test takers' ability, that also take less time to administer and rate than paper-and-pencil tests (Davidson and Coombe, 2012). Computerized testing can also have a significant positive impact on curriculum by providing test takers and teachers with immediate diagnostic feedback that can result in individual learning and teaching plans.

However, as Davidson and Coombe (2012) maintain, computerized testing is not used to any significant degree in education for reasons that are more theoretical than technical. The possibilities of computerized testing signal a need to undertake a major re-examination of the key principles of testing, namely validity, reliability, and practicality. Specifically, more research needs to be conducted to ensure that computerized tests are in fact measuring what they claim to be measuring (Davidson and Coombe, 2012)

The same is also true with Web-based language testing (WBLT). WBLT utilizes the Internet as a medium of test delivery. As Shin (2012) claims, compared to CBT (computer-based testing), WBLT has more beneficial factors in language testing practices and qualities. One is logistical flexibility. Another is that WBLT can enhance different aspects of authenticity: situational and interactional authenticity. Still another is that the reliability can be improved in WBLT because of consistent scoring by computer with regard to common construct features (Shin, 2012).

Shin (2012) stresses the efficiency of WBLT in language tests as follows: First, it enables test takers to take the test at their convenience. WBLT also provides a relatively simple and easy tool for language teachers who want to develop their own tests. Moreover, with the help of Natural Language Processing (NLP), in which test takers' lengthy

responses are analyzed and scored by algorithms constructed on the basis of pattern recognition, key word, and exact scoring, extended responses, such as speech samples, can be automatically scored. This function is extremely useful in speaking tests. In the WBLT context, speech samples have been scored by Speech Rater deployed for the TOEFL iBT Speaking Practice test. Although the ultimate goal of Speech Rater v 1.0 is to provide diagnostic feedback, it can now only provide learners with immediate score feedback. Another example using NLP in spoken language testing is Pearson's Versant in which the test taker responds orally to a series of recorded spoken prompts on various item types, including sentence repetition, building short-answer questions, and story retelling. In the Versant test, test takers' responses are scored automatically on the categories of sentence mastery, vocabulary, fluency and pronunciation, and listening comprehension (Shin, 2012).

However, there are drawbacks in the use of WBL, as Shin (2012) points out as follows: First, the fact that a test is delivered via the Internet may introduce some construct-irrelevant variance, which is often related to varying degrees of computer or Web familiarity among test takers. Second, security problems may be an issue. Third, despite the fact that automated scoring enables test takers to produce the constructed responses in large scale tests, the automatic scoring system itself can be another source of measurement error when it fails to take into account synonyms or paraphrases as alternative answers. Finally, technical devices can pose serious problems for high-stakes tests when test takers' responses may not be stored correctly (Shin, 2012).

Granted that those pit falls have been pointed out, with the help of advanced and network technology, WBLT has become a major medium of test delivery for both low-stakes and high-stakes tests. Online testing can improve the way we measure language ability in various ways. Particularly, WBLT may

improve test authenticity and reliability by making possible a rich contextualized input, various response formats, and automated scoring. Nevertheless, there has been little evidence to suggest whether online testing can actually help us to make more valid interpretations and uses of test scores. Therefore, more research should be conducted to see if reliable scoring can be obtained through automated scoring systems for constructed responses (Shin, 2012).

6. Conclusion

It is commonly believed that tests of spoken language ability are the most difficult to develop and administer. Despite a growing demand for speaking ability assessment, there remain a number of areas of great concern to language testers. Most notable of these concerns are the effect on performance of characteristics of the test taker and of the interlocutor, construct definition (what is it we are actually trying to test?), the predictability of task response, the validity, and the consistency of the scoring system (O'Sullivan 2012). People may still argue that speaking is the most difficult to test. However, it can be relatively straightforward if we are well prepared and are willing to accept that any test is a fine balance between all the competing requirements (cf. O'Sullivan, 2012). There should be a trade off among the necessary conditions in each testing context.

When it comes to an integrated assessment of speaking, as Plakans (2012) mentions, investigating and advancing constructs for integrated skill assessment is a promising and challenging field. There is a continued need for research as the concept of integrated skills assessment continues to evolve.

References

Brown, J. D. (2012). Choosing the right type of assessment. In Coombe, C., Davidson, P., O'Sullivan, B. and Stoyhoff, S.(Eds.), *The*

- Cambridge Guide to Second Language Assessment* (pp. 133-139). Cambridge: Cambridge University Press.
- Chapelle, C.A. (2013). (ed.). *The Encyclopedia of Applied Linguistics*. UK: Wiley-Blackwell.
- Coombe, C., Davidson, P., O'Sullivan, B. and Stoyhoff, S.(Eds.). (2012). *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Davidson, P. and Coombe, C. (2012). Computerized language assessment. In Coombe, C., Davidson, P., O'Sullivan, B. and Stoyhoff, S.(Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 267-273). Cambridge: Cambridge University Press.
- Katz, A. (2012). Linking assessment with instructional aims. In Coombe, C., Davidson, P., O'Sullivan, B. and Stoyhoff, S.(Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 66-73). Cambridge: Cambridge University Press.
- Lee, Y-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, 23, 2, 131-166.
- Henkel, E. (2010). Integrating the four skills: current and historical perspectives. In R.B. Kaplan (Ed.), *The Oxford Handbook of Applied Linguistics* Second Edition (pp.110-123). UK: Oxford University Press.
- (MEXT) The Ministry of Education, Culture, Sports, Science & Technology. (2012, February 29) Retrieved February 29, 2012 from http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/eiyaku/1298353.htm.
- O'Sullivan, B and Nakatsuhara, F. (2011). Quantifying Conversational Styles in Group Oral Test Discourse. In O'Sullivan, B. (Ed.). *Language Testing: Theories and Practices*. (chapter 8) New York: Palgrave Macmillan.
- O'Sullivan, B. (ed). (2011). *Language Testing: Theories and Practices*. Palgrave Advances.
- O'Sullivan, B. (2012). Assessing speaking. In Coombe, C., Davidson, P., O'Sullivan, B. and Stoyhoff, S. (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 234-246). Cambridge: Cambridge University Press.
- Plakans, L. (2011, September 5). Integrated Assessment. Retrieved September 5, 2011, from <http://languagetesting.info/video/main.html>
- Plakans, L. (2012). Writing integrated items. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of*

- Language Testing* (pp.249-260). UK: Routledge.
- Plakans, L. (2013). Assessment of Integrated Skills. In C. A. Chapell (Ed.), *The Encyclopedia of Applied Linguistics* Vol. I (pp.204-212). UK: Wiley-Blackwell.
- Shin, Sun-Y. (2012). Web-based language testing. In Coombe, C., Davidson, P., O'Sullivan, B. and Stoyhoff, S. (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 274-279). Cambridge: Cambridge University Press.
- Taylor, L. (2011). Introduction. In L. Taylor (ed.), *Examining Speaking: Research and practice in assessing second language speaking*. Cambridge, UK: Cambridge University Press.