Plenary II

Theory and Practice in Speaking Assessment

Yong-Won Lee
(Seoul National University)

ABSTRACT

In today's globalized World, oral communication competence in English is an essential skill for not only internationally-active scholars and researchers in academia but also people in business, industry, and many other sectors of our society. The main purpose of the paper is to give a brief overview of theories and practice of EFL speaking assessment, with a particular emphasis on the construct definition, task design, and rating scales. More specifically, the paper will: (a) first discuss the construct of speaking to be assessed, (b) then describe the major components of the assessment development process (including test/task design, rating scale development/validation, rater training), (c) and finally identify some new formats and technology of speaking assessment that deserve further exploration and investigation in the future.

Key words: Speaking assessment, construct definition, task design, rating scales

I. Introduction

In today's globalized World, oral communication competence in English is an essential skill for not only internationally-active scholars and researchers in academia but also people in business, industry, and many other sectors of our society. Nevertheless, the teaching of speaking has been neglected for many years in foreign language teaching, including English as a foreign language (EFL), for various reasons (e.g., the strong influence of the grammar translation method, lack of English native speaker teachers and EFL teachers with near-native English proficiency, large class size) (Nation, 2011). Increasing use of communicative language teaching methods in classroom settings for the past several decades have probably helped English teaching professionals to rediscover the value and importance of speaking in foreign language teaching and learning.

In relation to such trends, the development and planned administration of the National English Proficiency Tests (NEAT) have also re-generated interests recently among Korean EFL teaching professionals in the teaching and assessment of speaking in the Korean EFL context. All three levels of the NEAT exams (Levels 1, 2, and 3) include the speaking section as an essential component of the test. While the inclusion of the speaking section in each of these tests is expected to bring some positive washback impact on English education in Korea, it has posed a variety of challenges not only for language assessment specialists but also major testing institutions. Not to mention the scoring and rater training issues for the speaking sections of these large-scale assessments, it has become a very important job of assessment developers to clearly define the assessment constructs for these proficiency-oriented speaking tests and provide sufficient validity evidence for the tasks and rating scales developed to tap into these constructs.

With these backdrops, the main purpose of the paper is to give a brief overview of theories and practice of EFL speaking assessment, with a particular emphasis on the construct definition, task design, and rating scales. More specifically, the paper will: (a) first discuss the construct of speaking to be assessed, (b) then describe the major components of the assessment development process (including test/task design, rating scale development/validation, rater training), (c) and finally identify some new formats and technology of speaking assessment that deserve further exploration and investigation in the future.

II. The Construct of Speaking

1. Defining the Construct

A sound construct definition of what it means to be able to speak is crucial for the study of second language speaking assessment. Fulcher (2003) explains *speaking* in terms of the verbal use of language to communicate with others. In other words, unlike other skills of language use—namely, reading, writing, and listening—speaking consists of verbal as well as productive skills. Even though both written and spoken language involve productive skills, the linguistic features observed in speaking are different from those observed in writing. According to Bygate (2009), speaking consists of three main linguistic features: phonological, lexico-grammatical, and discouse features. The phonological features include segmental as well as supra-segmental features; the lexico-grammatical features deal with not only sentential but also with morphological level utterances; and discourse features involve socio-pragmatic features which are inherent to interaction.

Furthermore, these linguistic features that distinguish *speaking* from other language skills are controlled by the time-pressure and reciprocity conditions (Bygate, 2009). Given that speaking is a process that involves a two-way interaction done in real-time, participants in the interaction are subject to time constraints (Burns & Seidlhofer, 2002). Therefore, there are occurrences of various incomplete and short idea units, as well as hesitations, pauses, false starts, and repairs. Regarding the frequency of the words used in speaking, speakers tend to use less formal vocabulary and more coordination (rather than subordination). Additionally, in a conversation, pre-formulated expressions and idioms are repeatedly utilized. All of these features demonstrate the way speakers try to cope with the time-constraints. By the same token, reciprocity conditions influence the linguistic features observed during a conversation. For instance, repairs might occur when speakers think they need to make their utterances clearer, but they might also occur when, although clear to the speaker, the listener did not understand the utterances. In addition, the interactive characteristic of speech brings into play other types of features, such as turn-taking, initiation-response-feedback (IRF), and adjacency pairs.

Accordingly, the process of speaking consists of internal as well as external decisions. On the one hand, speakers have to make internal decisions as to what to say and how to say it. These are referred to as psycholinguistic decisions. On the other hand, speakers have to make external decisions concerning how to participate in an interaction by taking into consideration the utterances of the interlocutor. These are referred to as sociolinguistic decisions.

2. Theoretical Models of Speaking/Proficiency

Throughout several decades of research on linguistic competence, linguists have developed various models that attempt to explain the different components of linguistic competence. At the same time, these models have been used as a reference framework to define a learner's overall proficiency and the way it is measured (Luoma, 2004). For this reason, understanding the models is key to the process of defining the construct of speaking.

One of the most commonly discussed models of linguistic competence is that of Chomsky (1965). This model proposed a clear difference between linguistic *competence* and linguistic *performance*. While the former component explained the internal knowledge of language and its structure, the latter described the “processes of encoding and decoding” involved in the process of communicating with others. This differentiation, however, has been a matter of debate (Fulcher, 2003). Some believe that the linguistic performance is real, whereas the linguistic performance is incidental. Others have defined their constructs without any explicit distinction between them, arguing that communication strategies are simultaneously both internal and external.

Other models have focused more specifically on the interaction between both internal knowledge or competence of a language and the ability to use this competence to communicate with others. Canale and Swain (1980), for example, developed a model of communicative competence that consists of two major components: grammatical competence, which deals with “lexis, morphology, sentence-grammar semantics, and phonology”; and sociolinguistic competence, which includes “sociocultural rules and rules of discourse” (Bachman, 1990:85). Canale (1983) proposed that, apart `from these two major components, a further distinction should be made between sociolinguistic competence and discourse competence that included cohesion and coherence.

Building upon these models, Bachman (1990) proposed a new model of communicative language abilities that was supported by empirical evidence. This model consists of two major components: organizational competence and pragmatic competence. Whilst the former component deals with grammatical and textual competence, the latter is composed of illocutionary and sociolinguistic competence. In another model by Bachman (2010), language use is described as a process of interaction. The model can be illustrated by the interchange of language between a customer and a waiter. Both speakers bring their own topical and language knowledge, personal attributes, strategic competence, and cognitive strategies to the conversation.

More specific to speaking, Bygate (1987) developed a model that described the knowledge and skills that a person needs in order to speak. Within this model, speaking is considered an internal process that is composed of three major stages: planning, selection, and production. Each stage requires certain knowledge and a skill. For example, during the planning stage, a person needs knowledge of conventions (i.e., informational and interactional) and of the state of the discourse. At the same time the learner needs message planning skills and management skills (i.e., turn-taking skills).

Another model, proposed by Levelt (1989), includes four major phases—namely, conceptualization, formulation, articulation, and monitoring. This model is more specific when describing the different internal processes from preverbal messages to overt speech. Apart from production components, it also includes a speech comprehension component which, as mentioned earlier, is also a crucial part in a two-way interaction.

III. Development of Speaking Assessments

1. Purposes of Assessing Speaking

The most common purpose of assessing speaking is associated with the positive washback effect on foreign language learning. It has been argued that assessing speaking could encourage learners to actively learn and practice speaking in classrooms (Swain, 2001), which would lead to a more communicative approach to learning a language. Nevertheless, others argue that the real washback effect comes from the “adequate informative feedback” drawn from the results of the speaking assessment (Fulcher, 2003: 173). In other words, the results of an oral test could provide instructors, as well as learners, with diagnostic information of which language skills have already been acquired or are yet to be acquired by the learner. This information can be used to personalize classroom curriculums by targeting the learners’ weakest language skills.

Furthermore, another use of speaking assessments includes the screening of individuals for educational or commercial purposes. This is common in several educational programs that require students to have a certain level of fluency in English speaking. In this case, English speaking tests might be useful in screening out those students who do not match the requirements for a particular program. The third important use of speaking assessments, which is closely related to screening, is selection. Particularly in Korea, this is a widespread use of tests. Employers use speaking tests to select the most apt employees for their companies. Finally, another purpose of assessing speaking could be placement. This refers to the use of tests to assign learners to a certain instruction level that best fits with their educational needs.

2. Designing and Developing Speaking Assessment

One of the major concerns in the process of designing and developing speaking assessment tasks is whether they directly relate to the speaking constructs. Since tasks are the tools for observing the learners' language abilities, they should try to elicit language that reflects what is intended to be measured. Accordingly, variables such as task type, topic, register, and conditions, among others, should be carefully taken into consideration when we develop tasks. In the case of topic, for example, previous experience or background knowledge might directly influence the difficulty level of the task. Furthermore, the register (i.e., formal or informal) of the language elicited might also be a factor that affects the speaking performance. In terms of the task conditions, test-takers' performances might fluctuate according to time constraints or whether preparation time is given or not.

Several studies have suggested different ways to describe tasks. Weir (1993:39) describes tasks in terms of performance conditions that include time constraints, degree of reciprocity, purpose, interlocutors, setting, role, topic, channel, input dimensions, size, complexity, and range. Fulcher (2003:57) provides a more concise framework for describing tasks which consists of task orientation, interactional relationship, goal orientation, interlocutor status and familiarity, topic, and situations. Notwithstanding the differences among frameworks, their importance lies in the relationship between these different factors and the construct of the speaking test. If, for instance, a speaking test intends to assess the interactive ability of test-takers in a business situation, the tasks should include a two-way interactional relationship and the topic should be business-related.

There are several formats of speaking assessments which include semi-direct tests, interviews, paired, or group tests. Semi-direct tests are characterized by one-way interactions between the test-takers and a computer. The tasks elicit a monologic response to a wide variety of questions involving reading sentences, telling a story related to a picture-sequence, or telling directions based on a map. The advantage of this type of test is its practicality, since a great number of test-takers can be tested at once. Nevertheless, a major drawback is the lack of tasks involving two-way or multi-way interactions. Furthermore, this type of task is less well-researched, compared to other speaking tasks.

Interviews are the most studied type of speaking tests. Unlike semi-direct tests, they are considered more authentic since they include tasks with two-way discussions. Some of the tasks used in interviews include role-plays, picture descriptions, and responses to open ended questions. While this type of test might allow test-takers to interact and engage in a conversation, there is still skepticism regarding the authenticity of the interactions. The asymmetric relationship between an interviewer and a test-taker is one of the major concerns when dealing with this format of speaking assessment. Several studies (e.g., Ross, 1992; Ross & Berwick, 1992; Brown, 2003) have demonstrated that the interviewers appear to have control over the topic of the conversations. As a response to these concerns, paired and group assessments have recently been the focus of interest in the language testing community. Interactions between test-takers have proven to be more symmetrical and to have richer speech functions (Taylor, 2001).

3. Scoring Rubrics

Besides the careful design of tasks, developing scoring rubrics is another crucial element in speaking assessments. Scoring rubrics should be designed in such a way that they accurately reflect the different characteristics of the speech samples collected during the tests. The design of scoring rubrics might differ depending on who uses them; that is, whether they are user-oriented, assessor-oriented, or construct-oriented (McNamara, 1996). A scoring rubric that is user-oriented would likely contain information that could be helpful for test-takers to improve their speaking performance, whereas a construct-oriented rubric might provide information to the test developers on how to improve a certain task.

Furthermore, there are two major types of scoring rubrics—holistic and analytic (Fulcher, 2003, Luoma, 2004). The former refers to an overall score of the test-taker's language performance, while the latter consists of several subscales that may include grammar, vocabulary, pronunciation, fluency, and discourse management among others. Although the analytic scales might seem finer-grained when compared to holistic scales, they still reflect an overall score for each of the analytic rating dimensions. Future research should focus on ways to reflect more specifically various characteristics in each of the analytic criteria.

Throughout the history of speaking assessment, rating scales have undergone several changes. During the early 1950s, Foreign Service Institute (FSI) rating scales were used to score speaking performances (Fucher, 2003). This scale consists of five-score points and five criteria: accent, grammar, vocabulary, fluency, and comprehension. Due to the lack of representation of the lower-level speech samples within the levels described in this scale, in 1968 the Interagency Language Roundtable (IRL) Rating scale was developed. This rating scale consists of 6 bands and includes the same analytic criteria as the FSI scales. Later on, in 1986, the American Council on the Teaching of Foreign Languages (ACTFL) rating scale was introduced. This scale includes levels such as Novice (subcategorized as low, mid, high), Intermediate (subcategorized as low, mid, high), Advanced, Advance Plus, and Superior. In 2001, the Common European Framework of Reference (CEFR) rating scale was developed. It consists of 5 analytic scoring dimensions, which include range, accuracy, fluency, interaction, and coherence. This scale is characterized by a 6-score band, namely Basic user (A1, A2, A2+), Independent user (B1, B1+, B2, B2+), and Proficient user (C1, C2).

IV. Future of Speaking Assessment

1. Paired/Group Assessment

As mentioned above, in an attempt to cope with some of the limitations of the interview format of speaking assessment, paired and group assessments have been developed. These tests have the advantage of assessing test-takers' interactive language skills with a more authentic approach. Nevertheless, there is a lack of empirical studies in support of the construct validity of these tests. Some argue that the interlocutor's personality, age, gender, and language proficiency might directly affect the test-takers' speaking performance. Moreover, future studies should focus on scoring rubrics that could precisely represent the types of speech elicited with these types of tests.

2. Automated Speech Scoring

Certainly, there is an urgent need to find more practical ways to score speaking tests. Automated speech scoring and feedback systems have been suggested as a viable option. However, research in this particular area is scarce. More empirical evidence should be presented to demonstrate the relationships between the scores from automated systems and the human raters and the meaning of the automated scores.

V. Concluding Remarks

The ultimate goal of speaking assessment is to make inferences about the test takers' speaking proficiency based on their speaking scores, or performance on speaking tasks, and to make important decisions about the test-takers based on such assessment data. In order to defend the decisions made about the test-takers based on these assessment results, assessment developers and validators should provide sufficient evidence in support of the validity and fairness of their assessment outcomes. This means that all of the major steps involved in the process of assessment development and use should be evaluated in an integrated way, which may include construct definition, creating test specification, creating tasks and rating scales, scoring, score interpretation, decision making, and so forth. In this sense, the whole process of assessment development and use should be examined as an interconnected chain of actions for the purpose of building validity arguments for a particular speaking assessment of interest.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language Assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20, 1–25.

Burns, A., & Seidlhofer, B. (2002). Speaking and pronunciation. In N. Schmitt (ed.), *An introduction to applied linguistics* (pp. 211-232). London, UK: Hodder Education.

Bygate, M. (2009). Teaching and testing speaking. In M. H. Long & C. J. Doughty (eds.), *The handbook of language teaching* (pp. 412-440). Malden, MA: Blackwell-Wiley.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (eds.), *Language and communication* (p. 2-27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, *1*, 1-47.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, Massachusetts: The M.I.T. Press.

Fulcher, G. (2003). *Testing second language speaking.* New York, NY: Pearson/Longman.

Levelt, W. (1989). *Speaking: from intention to articulation.* Cambridge: MIT Press.

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

McNamara T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.

Nation, I. P. (2011). Second language speaking. In E. Hinkel (ed.), Handbook *of research in second language teaching and learning* (pp. 444-454). New York, NY: Routledge.

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language testing, 9*, 173-186.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in second language acquisition*, *14*, 159-176.

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 275-302.

Taylor, L. (2001). The paired speaking test format: recent studies. *Research Notes*, 6, 15-17.

Underhill, N.(1987). *Testing spoken language*. Cambridge, UK: Cambridge University Press.

Weir, C. (1993). *Understanding and developing language tests*. New York, NY: Prentice Hall.

About the Presenter

Dr. Lee received an MA in English Education from Korea National University of Education, Korea and a Ph.D. in Speech Communication from the Pennsylvania State University, USA, specializing in ESL/EFL assessment. He has worked as a research scientist in TOEFL/ELL research group at Educational Testing Service, USA for more than 8 years, and since 2007 he has been teaching applied linguistics and language assessment courses at Seoul National University. His professional interests include language assessment, language planning/policy, sociolinguistics, and English as a glocal language. Currently, he is an associate professor at the Department of English Language and Literature, Seoul National University. Email: ylee01@snu.ac.kr.