〈原著論文〉

Working from the Construct Definition: Assessing University Speaking Classes

Derek N. CANNING

Abstract

This paper looks at common conceptions of the speaking construct and how this has informed research into speaking assessment. A great deal of research has been done on the construct definition of speaking and how it can be applied to high-stakes, norm-referenced testing. However, comparatively little has been written on the multi-componential nature of speaking and classroom assessment. University-level classroom assessment often relies on paper-based cloze and listening tests with low face validity. This paper proposes classroom assessment methods that focus on the range of skills identified by the speaking construct definition. These assessment methods gauge communicative competencies such as establishing turns at talk and adjacency pairs, as well as linguistic competencies including grammar and fluency.

Key words: Speaking, Assessment, Construct Definition, Testing, Oral Communication

Introduction

There are gradations in ability from first being able to express yourself orally in a language to high-level proficiency. However, there is no clear consensus on how to classify these gradations. Gauging speaking performance solely by grammatical accuracy ignores other, equally vital components of speaking, including strategic and communicative competencies. Assessment methods such as the American Council on the Teaching of Foreign Languages (ACTFL) and the Interagency Language Roundtable Oral Interview (IRL) have attempted to assess speakers according to what they can do in the target language, though these have been criticized for a measurement scale of speaking ability that ranges from "zero" language proficiency to "perfect" proficiency (Bachman & Savignon, 1986). Bachman and Savignon argue

基礎総合教育部
論文受理日 2018 年 11 月 13 日

further that comparing second language speaking against the standard or norms of native performance is problematic in that there is no one example of native speech against which all other utterances are judged.

Other problems with tests such as the ACTFL and the ILR are, firstly, the "imprecision of descriptive criteria," based as they are on experience, rather than "systematic empirical or theoretical inquiry" (Cumming, 2009, p. 92). Bachman and Savignon share this concern. Secondly, the level bands may not be appropriate for different learners in variable contexts. Bachman and Savignon recommend considering the needs and goals of the learners when designing assessment tools. They further recommend rating scales that are context-independent and tests that can be tailored to the candidate's interests and needs. Their suggested rating scale evaluates a candidate's command of grammar, cohesion, and register. Three broad levels are suggested, from exhibiting no control over the stated characteristics to complete, error-free control (Bachman & Savignon, 1986).

The goal of many entry-level speaking courses in Japanese universities has shifted from traditional, grammar-centered instruction to a focus on a "communicative paradigm" (Matsuura, Chiba, and Hilderbrandt, 2001, p. 79). Classroom activities and assessment methods in speaking classes must reflect teaching methodologies that stress communicative learning. The criteria suggested by Bachman and Savignon for assessing proficiency in speaking are, however, too broad to capture subtle differences in student ability and the incremental progress university students might make over one or two semesters. Classroom speaking assessments can be improved upon by, in the first place, not judging performance against native speaker proficiency, and secondly, by grounding the language of the assessment tools in the interests and needs of the speakers themselves. Finally, assessment tools can be improved by identifying the sub-skills, or components, of speaking, and assessing those individually.

As complex as speaking is, it can be broken down into its components. Defining the construct of speaking makes clear to learners and instructors what criteria make up proficiency in speaking performance, thus improving the face validity of course assessments (Nation, 2009 and Fulcher, 2003). Identifying and understanding these components can then guide assessment techniques. These criteria, from grammatical accuracy to phonology and communicative competence, can then be used to inform course design and student assessment. This paper will look at construct definitions of speaking, how these have been used in performance assessment, and how they can be applied practically in a university-level speaking course.

Literature Review

A great deal of research has been done on the assessment of speaking in a second language and its relationship to the construct of speaking. At the most basic level, scoring a test involves deciding two problems: "what to count and how many points to assign to each component that you are counting" (Folse, 2006, p. 226). This literature review will look at how researchers have explored the first of these problems: what to count, or what is the construct definition of speaking.

In 1965 Chomsky highlighted the distinction between language competence and performance. 'Competence,' in Chomsky's definition refers to a speaker's knowledge of the language and 'performance' describes the use of language in speaking. In 1980, Canale and Swain developed the idea further and argued that in assessing competence, it was necessary to consider not only grammatical competence but a speaker's sociolinguistic competence, or their knowledge of the "rules of language use" (p. 6). Crucially, Canale and Swain noted that only language performance is observable. A speaker's competencies, or the actual scope of their knowledge of the language, can only be inferred from assessing a speaker in actual communicative performance. For that reason, they argued that syllabus design and language assessment should incorporate authentic communicative situations.

Fulcher (2003) notes that the distinction between competence and performance, or what he terms "the internal" and "the external" is fluid. When defining the speaking construct, Fulcher argues, it is not always necessary to distinguish between 'internal' competencies such as grammar or lexis and 'external' competencies such as strategy use or turn taking. A construct definition will incorporate both, and the tools of assessment allow us to examine the discrete components of that construct. Fulcher's 2003 book, by synthesizing language competencies and performance into a comprehensive construct definition, provides a starting point for the examination of the first challenge: what to count. Table 1 is a partial list of the components of speaking identified by Fulcher.

Fulcher does not claim that the criteria partially reproduced in Table 1 is an exhaustive picture of the construct of speaking. He also cautions that valid speaking assessments do not need to test a candidate's facility in each of the criteria individually. The construct definition outlined here is a menu of sorts. Criteria can be selected individually or combined in more complex assessments. The challenge then is to identify the relative importance of the chosen

Table 1　*Fulcher's speaking construct framework*

Language competence
● Phonology
● Accuracy
● Fluency
Strategic capacity
Textual knowledge
● The structure of talk
Pragmatic knowledge
● Appropriacy
Sociolinguistic knowledge

Note. Adapted from Fulcher (2003).

components to the overall construct.

Recent studies have attempted to identify the importance of speaking components quantitatively. A 2012 paper by De Jong, Steinel, Florijn, Schoonen, and Hulstijn breaks down these components into declarative knowledge (vocabulary and grammar) processing knowledge quickly (lexical retrieval), and pronunciation (quality of speech sounds, word stress, and intonation) and found that most were "...strongly associated with speaking proficiency" (2012, p. 29). Another study by Iwashita, Brown, McNamara, and O'Hagan (2008) examined the components of speaking that had the largest effect on raters' score of TOEFL ibt and found that raters attended to the same broad categorizations of speaking proficiency, including vocabulary, fluency, grammatical accuracy, and pronunciation. The Iwashita and De Jong studies both support the legitimacy of a construct definition like Fulcher's. The Iwashita paper notes that any one component of speaking does not account for a global assessment of competence. How a listener assesses competence is an "on balance" judgment made based on criteria even broader than the construct definition defined thus far.

Raters can grade speaking exam scores as a holistic piece, or as a composite of individual scores of skills and extra-componential criteria. Studies have looked at how raters who assign holistic scores arrive at their final assessment. Bøhn (2015) used think-aloud protocols to have teachers identify the salient sub-skills they attended to when grading the speaking performance on a Norwegian high school exam. Préfortaine and Kormos (2016) carried out a similar study in which raters were asked to identify the components assessed in their assignment of a global fluency score. Both found that raters attend to different components of speaking when performing assessments. Nevertheless, within each study there were broad agreements among the raters as to the importance of key components, such as linguistic competence in the case of

the Bøhn study and rhythm and effortlessness in the Préfortaine and Kormos study. The components attended to by the raters in these studies conform to the construct definition given by Fulcher.

The Bøhn study also found that raters attended to different rating criteria depending on the level of the test taker. Raters identified linguistic criteria as more salient for lower level speakers and factored the content of the examinees' speech into the holistic scores of higher-level speakers. The idea that the construct definition of speaking changes with the level of the speaker is a theme that runs through much of the research. A 1993 study by Pollitt and Murray found, like the Bøhn and Préfontaine and Kormos studies, that what aspects of speaking raters regard as relevant to proficiency changes with the level of the speaker. Lower-level speakers are evaluated to a greater degree on their grammatical competence. The importance of sociolinguistic and discourse competence is given more weight at higher levels of speaking.

Also noted in the Bøhn study as well as in the Pollitt and Murray paper is the effect of criteria not associated with the construct of speaking, such as the effort shown on the part of the candidate or his or her personality. These and other paralinguistic and construct-irrelevant criteria are found by other researchers as well. Studies by Ang-Aw and Goh (2011), May (2006), and Han (2016) all make explicit mention of the effect of non-criterion factors on rater scores. Rater effect is a separate avenue of study. It is enough here to simply acknowledge that factors not included in the construct definition play a role in the assessment of speaking.

In contrast to studies that looked at tests without set scoring criteria, other studies have looked at how raters conceptualize the speaking construct when given an operationalized rubric for scoring speaking assessments. Ang-Aw and Goh (2011) found discrepancies in rater understanding of the criteria used to give an overall speaking score in a high-stakes Singaporean English test. While raters may have given the same candidate similar scores, qualitative analysis showed that the reasoning behind the scores were very different. A study by Sawaki (2007) found a high degree of correlation among discrete components of assessment, including grammar, organization, cohesion, vocabulary, and pronunciation. The careful design of the latter test may account for its greater validity. What is significant about these two studies is that the more explicit rubric in the Sawaki study concentrates the discourse around the construct definition. In the Ang-Aw study, a lack of clarity, training, or both means that rater effect becomes more significant.

The picture that emerges from an overview of recent studies is that testing speaking by assessing clearly defined criteria is a valid approach. However, there is no clear consensus on

the relative importance of individual criteria to the overall construct of speaking. One clear pattern is the tendency for raters to assign greater importance to grammatical competence in lower-level speakers. Another is the effect of construct-irrelevant variance. There are three clear implications for designing assessment tools for university level speaking classes. First, assessments need to multi-componential. Second, there must be a focus on lower-order linguistic skills, such as grammar and fluency for lower-level students. Finally, construct-irrelevant variance must be mitigated through clearly defined scoring rubrics.

Assessment Criteria

This section will outline the criteria that can be selected for evaluating students' speaking performance, the rating scales used to score those criteria, and the methods by which they are assessed. Students can be assessed in two contexts: in paired conversation and in an interview with the instructor. Both are discussed in turn below.

Paired conversations

Assessing paired conversation lends opportunities to observe students in realistic scenarios. This fulfills one of Canale and Swain's (1980) fundamental prescriptions for communicative language teaching and assessment, that is, that performance be evaluated in authentic communicative situations. To observe and evaluate the candidates' use of communicative strategies, it is necessary to "mimic real-life encounters as much as possible." (Caban, 2003) Dyadic communication also allows the rater to observe and evaluate "features of interactional competence" such as turn taking, interactional listening comprehension, and non-verbal communication (Han, 2016).

In the assessment, students engage in an activity called "Peer Talk," a two-minute conversation with a partner randomly assigned from among the others in the class. The format will be familiar to the candidates, as this activity can be a regular exercise in speaking class. Candidates engage in a conversation on any conversational topic for two minutes. The rater will observe and time the conversation without participating. The scoring rubric is shown below in Table 2.

Several of the criteria listed above can be regarded as reflections of the candidate's sociolinguistic competence. The criteria "used/responded to opening" and "used/responded to closing" are intended as a means of assessing the candidates' basic competency in turn taking

Table 2 *Peer-Talk scoring rubric*

Used/responded to opening	Yes	Partially	No
Asked (minimum) one appropriate question	Yes	Partially	No
Asked (minimum) one follow-up question	Yes	Partially	No
Reacted naturally to conversation	Yes	Partially	No
Filled pauses	Yes	Partially	No
Used natural pronunciation	Yes	Partially	No
Offered/sought clarification	Yes	Partially	No
Used/responded to closing	Yes	Partially	No

and adjacency pairs. The criteria "asked one question" and "asked one follow-up question" are a very simple means of evaluating the candidate's locutionary listening comprehension and his or her ability to interact with his/her interlocutor. The appropriateness of the question can be determined by how natural the question is in daily conversation. For example, "What are you doing after this?" is a more natural question than, "What's your favorite color?" Rater judgment is a factor here. The criterion "reacted naturally to conversation" is again dependent on rater judgment but is another metric of sociolinguistic competence. Appropriate reactions can involve non-verbal communication, verbal reactions to unexpected, upsetting, or funny news, or follow-up questions that indicate the listener has understood his/her interlocutor's utterances.

Offering or soliciting clarifications can be taken as indicative of the candidate's proficiency in strategic competence, at least as it is defined by Canale and Swain, that is, as "the strategies that may be called into action to compensate for breakdowns in communication" (1980, P. 30). The criteria, "asked a follow-up question" and "reacted naturally to conversation" can also be regarded as a metrics of strategic competence. The boundary between sociolinguistic competence and strategic competence is not always clear cut.

Significantly, except for phonology and one metric of fluency, language competence is not assessed in the paired conversation test. Fulcher (2003) notes that assessing fluency is problematic, as it can be difficult to determine a speaker's purpose in pausing while speaking. Despite this caveat, it is possible to use a simple metric such as noting pauses as either filled or unfilled. Pronunciation is graded simply on its "naturalness," an admittedly vague concept that leaves a great deal to rater judgment. Natural pronunciation can be regarded as pronunciation that does not interfere with the meaning of utterances. Elements of grammatical competence are evaluated in the second, interview phase of the assessment.

Interviews

The interview assessment affords an opportunity to more closely evaluate the candidate's grammatical competence, command of phonology, and a metric of fluency not assessed in the paired speaking phase. Additionally, the interview test assesses the candidate's comprehension of and facility with the course material covered in class with the text book.

In this assessment, candidates speak one-on-one with the instructor, who asks 10 questions adapted from the sequenced functions in the text book. Each reply is evaluated individually. The rubric for evaluation is given in Table 3 below.

The rubric is brief, allowing for the rater to attend to each reply in detail. The criterion, "responded appropriately to question" is a binary judgment; did the candidate answer the question logically or not. The criterion "length of reply" is intended as another evaluation of fluency, supplementing the assessment of pause phenomena done in the conversation phase. Length of reply can be evaluated as: insufficient, as in single-word answers or incomplete sentences; sufficient, or answers given in complete sentences; and detailed, or answers that include extra or expository information. "Grammatical accuracy" can be evaluated in one of two ways. A detailed, quantifiable assessment would involve the rater counting errors per clause. If this proves impractical, the rater can evaluate accuracy holistically as: low (grammatical errors interfere with meaning/understanding), adequate (grammatical errors are made but do not interfere with meaning), and excellent (grammatical errors are insignificant or do not occur).

Taken together, the criteria evaluated in the conversation test and the interview test are intended to assess, in a practical way, language competence, strategic capacity, textual, pragmatic, and sociolinguistic knowledge. Again, these are the key components of the construct of speaking, as presented by Fulcher (2003). The rubrics outlined here are not exhaustive. Furthermore, the test-tasks themselves are not likely to elicit a volume of speech that is adequate to reliably evaluate a candidate's speaking competence. Nevertheless, the inclusion of two distinct stages of testing, both of which require the production of speech in authentic situations, is a step towards a more communicative method of assessment.

Table 3　*Interview scoring rubric*

Responded appropriately to question	Yes		No
Length of reply	Detailed	Sufficient	Insufficient
Grammatical accuracy	Excellent	Adequate	Low

Conclusion

Many of the components of speaking in a second language are intuitively obvious and have been studied extensively. These include fluency, pronunciation, and command of the grammatical system. When evaluating speaking in a second language, these are often included as evaluation criteria. It was not until the 1980s, however, and Canale and Swain's insight that competency in speaking also entailed competency in the sociolinguistic system of the language. They would also argue that speaking further entails strategic competencies, or the methods a speaker employs when communication begins to fail. These insights have been incorporated into communicative teaching methods that involve speakers in real-life, real-time situations.

The value of communicative teaching methods is rarely contested. However, there is yet no consensus on how criteria as varied as grammar and strategy use can be validly, reliably and practically assessed in classroom situations. Instructor intuition in these matters is practical, but unreliable, dependent as it is on rater bias. Objective measurements of some components of speaking such as the speed of lexical retrieval, length of pauses, and errors per T-section have been shown to be highly reliable. For classes of up to 30 students, however, these types of assessments are impractical.

The solution is not to shy away from the complexity of speaking. A first step is to clearly define the components of speaking. Following that, assessment methods can be designed that can then be used to inform the content and methods of the course. This paper has shown how this process can be begun in speaking courses at the university level.

References

Ang-Aw, H. T., & Goh, C. C. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal, 42*(1), 31–51.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *The Modern Language Journal, 70*(4), 380–390.

Bøhn, H. (2015). Assessing spoken EFL without a common rating scale: Norwegian EFL teachers' conceptions of construct. *SAGE Open, 5*(4).

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*, 1–44.

Canale, M. & Swain, M. (1981): Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Chomsky, N., (1965). *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press.

Cumming, A. (2009). Language assessment in education: Tests, curricula, and teaching. *Annual Review of Applied Linguistics, 29*, 90–100.

De Jong, N., Steinel, M., Florijn, A., Schoonen, R., & Hulstijn, J. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5–34.

Folse, K. S. (2006). *The art of teaching speaking: Research and pedagogy for the ESL/EFL classroom*. Ann Arbor: University of Michigan Press.

Fulcher, G. (2003). *Testing Second Language Speaking*. Harlow: Pearson Longman.

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 16*(1). 1–24.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49.

Matsuura, H., Chiba, R., & Hilderbrandt, P. (2001) Beliefs about learning and teaching communicative English in Japan. *JALT Journal, 23*(1), 69–89.

May, L. A. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing, 11*, 29–51.

Nation, I. S. P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. New York: Routledge.

Pollitt, A. & Murray, N. L. 1996: What raters really pay attention to. In Milanovic, M. and Saville, N. (Eds.), *Performance testing, cognition and assessment*. Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press, 74–91.

Préfontaine, Y. & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching, 54*(2), 151–169.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*, 355–390.

英語スピーキング能力の定義から考える
――大学の授業の評価方法――

キャニング , D. N

抄　録

本稿では，英語スピーキング能力の一般的概念と，それがどのようにスピーキング能力の評価に影響しているかを調査する。これまでの研究では，スピーキングの定義や，それが TOEFL や TOEIC などの学力基準試験にどのように応用されてきたかが論議されている。しかしスピーキングの多角的要素やそれらの能力をどのように授業で測るかは，ほとんど研究がされていない。大学のスピーキングの授業評価は，主に妥当性が低いとされる穴埋めやリスニングテストが大部分を占めてきた。本研究では，スピーキングの定義や概念で示された様々なスキルを測る評価方法を提案する。この方法は，ペアによる会話のやりとり，そして文法や流暢さなどの言語運用能力を含むコミュニケーション力を測るものである。

キーワード：スピーキング，評価，言語力の定義，テスト，オーラルコミュニケーション