English Language & Literature Teaching, Vol. 21, No. 1 Spring 2015

A Study of Rating Criteria for Speaking Assessment: Using Empirical Data of EFL Learners*

Hera Chu
(Sun Moon University)

Chu, Hera (2015). A study of rating criteria for speaking assessment: using empirical data of EFL learners. *English Language & Literature Teaching*, *21*(1), 75-96.

The purpose of this study is to help establish an efficient speaking assessment tool for learners of English as a Foreign Language (EFL) by examining a degree of relevance of rating criteria employed and raters' perceptions of criteria for speaking. This study analyzed 11 native speaker raters' assessment on speaking performances of 22 Korean secondary school students participating in an English speech competition. Ratings were conducted based on the following 6 criteria: *presentation, content, memorization/timing, fluency, grammar,* and *pronunciation*. Findings exhibit first, relatively high correlations were found between the criteria and the final scores of students' performances except *pronunciation*. The most influential criteria in contributing to the students' overall speaking ability, however, were *fluency* and *content*. Also, there was a correlation between the final scores and scores the students obtained on a TOSEL proficiency test, adding improved validity of the criteria. Finally, raters' perceptions tend to be reflected in their ratings as the results of the questionnaire show *fluency, content* and *grammar, pronunciation* were perceived as important features. The results from the two analyses suggest that the criteria employed here are considered valid, while *fluency* and *content* may be more significant in determining learners' overall speaking ability. Finally, *pronunciation* did not turn out to be a strong indicator in any of the student performances. The reason for this could be explained in part by lack of detailed and explicit standards for *pronunciation*.

[**speaking assessment/rating criteria/correlation/fluency/pronunciation**]

* This paper was supported by the Sun Moon University research grant of 2012.

I. INTRODUCTION

Speaking is considered one of the most direct and basic means of expressing oneself. Especially with the emphasis of communicative competence, the significance of improving speaking skills for authentic communication has received particular attention in a Korean EFL context as well (Korea Institute of Curriculum and Education (KICE), 2005). As a consequence, English learners have been encouraged to practice and develop speaking skills using methods (i.e. Reading aloud, Interaction tasks, Oral presentation, Information transfer, Interview etc.) derived from the theory of communicative competence within or outside a classroom, but it appears that speaking is not still assessed at all or if assessed, it is not done adequately. (Kim, 2003; Luoma, 2004; Prapphal, 2008). In addition, even though the curriculum focus of English education is on the development of oral skills (KICE, 2005), students' oral skills are not assessed in University Entrance Exams, which in turn seems to lead to weakening effect on enhancing speaking skills.

There might be possible reasons why speaking is not assessed as efficiently or extensively as other language skills. Researchers (Eckes, 2008; Knight, 1992; McNamara, 2000; Hughes, 2003) point out difficulties in assessing speaking skills as follows: 1) assessing speaking skills is demanding and time consuming as it requires evaluating and producing results, 2) it is difficult to design productive and relevant speaking tasks to elicit student response, 3) selecting criteria is not easy because features defining speaking ability may differ depending on raters and the type of speaking task and so on, and 4) it necessarily involves subjectivity. To keep the level of subjectivity to its minimum, rater training is necessary for consistent rating for speaking. Kim (2003), in her study on the task type used for speaking assessment, reports other practical constraints in conducting speaking assessment in the secondary level of the Korean context as large class and heavy workload and so on. However, most noticeable among them includes teachers' lack of confidence in conducting speaking assessment as not many teachers are equipped with the theory of speaking assessment or had training in conducting speaking assessment (Kim, 2003). Furthermore, lack of efficient assessment instrument to use was noted as a problem, apart from the guidelines or rubrics individual teachers can gain on the Internet.

Thus, it is considered necessary to find a way to help teachers or practitioners conduct a more valid and reliable evaluation of speaking in a school setting. Yet, in practice, few studies have been conducted on assessment of speaking performances using empirical data in Korea. In particular, what is more lacking has been on the ways of assessing speaking performances using rating criteria and how to use them from a teacher's perspective. In order to help teachers assess their students' speaking more easily and

effectively, this study aims to identify how the each criterion is rated by native-speaker teachers to contribute to overall speaking ability and to investigate how their perceptions of rating criteria are reflected in their rating. Another aim is by showing how the criteria are rated, to make teachers more aware of the flexible nature of selecting and designing criteria depending on the context they are to be used. With these purposes in mind, this study attempts to answer the following research questions.

1) How does each of the rating criteria relate to each other to contribute to the final score of students' performance?
2) Is there any correlation between the criteria and the score of listening and speaking section on the TOSEL proficiency test? If any, how does each criterion, or relations between the criteria, contribute to the TOSEL score?
3) How do native raters perceive the relative importance of each rating criterion?

II. THEORETICAL BACKGROUND

1. Rating Criteria for Speaking Assessment

Speaking is believed to be one of the most difficult language skills to measure (Luoma, 2004; McNamara, 2000; O' Sullivan, 2008), although it is widely acknowledged by a backwash effect that the best way to promote a certain skill is to evaluate it. If developing speaking skills is essential, then speaking should be assessed. As pointed out, the primary concern for assessing speaking seems to lie in a difficulty designing or selecting efficient speaking assessment instruments as the outcome to be obtained from the assessment for speaking is affected by a variety of factors including definition of the nature of the speaking construct, the characteristics of the test-takers, the effect of task types, interlocutor effect, and rater reliability and so on (Brown, 2004, 2007; McNamara, 1996; O'Sullivan, 2008). Therefore, it is important to understand how these factors interrelated with each other to create an overall speaking score.

In developing or selecting assessment criteria for speaking, what should be done first is defining speaking, in other words, understanding what features construct a speaking skill since the features or sub skills constructing speaking, by being grouped together or as a single skill become a basis for creating rating criteria for speaking. There exist a variety of criteria used for speaking assessment, and selection of criteria can be decided depending on the context where the criteria are to be used. However, actual criteria considered essential in assessing a speaking skill are quite similar in content regardless of context variables such as a task type (Brown, 2007). Brown's (2007) list of possible

criteria for speaking succinctly illustrates the common features of criteria as follows: pronunciation, fluency, vocabulary, grammar, discourse features (cohesion, sociolinguistic appropriateness etc.), and task (accomplishing the objective of the task is related to comprehension). Knight's (1992) list of criteria for assessing speaking is also considered comprehensive. The features included are grammar, vocabulary, pronunciation, fluency, conversational skills, non-verbal skills, and content. Among the criteria listed in the two scholars' summaries, seven criteria most relevant to the criteria used in this study are summarized as follows:

- Fluency – an ability to speak at a normal speed without too much hesitation and flow naturally not worrying too much about being 100 % accurate.
- Grammar – an ability to use grammatical structure accurately
- Pronunciation – pronouncing individual sounds, applying word and sentence stress and rhythm, intonation, aspects of connected speech such as linking, elision, and assimilation
- Content – coherence and relevance of arguments, organization and topic development
- Non-verbal skills – the way we use body language such as posture, eye contact, gesture and facial expressions
- Sociolinguistic skills – the ability to initiate discourse, take turns when appropriate, keep the conversation going, use right vocabulary in the right context

2. Previous Studies on Speaking Assessment

A number of studies investigating rating criteria in order to help develop rating scales or to validate its use have been carried out using actual performances of students (Iwashita, McNamara, & O'Hegan, 2008; Jin & Mak, 2012; Plough, Briggs, & Van Bonn, 2010: Tajeddin, Alemi, & Pashmforoosh, 2011). Plough et al (2010), for example, examined the evaluation criteria used to assess oral performance of prospective graduate student instructors in two tasks- transactional and interactional in nature. The data analyzed by the rating criteria was then compared to the final scores the students obtained from a holistic rating scale and written comments given by the raters. Their findings are informative. The three analyses revealed corresponding results that pronunciation and listening comprehension were indicated as the two most detrimental features in describing the speaking proficiency of the students among others: vocabulary, grammar, transactional competence, and interactional competence. Pronunciation in their study was defined as a criterion which involves fluency and intelligibility.

Similar results were reported in the study by Tajeddin, Alemi and Pashmforoosh (2011) on the effect of a rater training program for non-native raters. Tajeddin et al. (2011)

examined the criteria the raters used before and after a rater training program. The raters first specified 10 criteria they used when assessing 10 students' monologs. After the training, changes in the importance they assigned to the criteria specified were observed. The non-native raters placed rising importance on fluency, comprehension, and organization while affective variables and the rate of speech were assigned less importance. Advocating a need for a rater training program, Tajeddin et al. (2011) concluded by suggesting that the rating program indeed provided the raters with a chance to "attend to macro-level, higher-components when making a global judgment of speaking performance" (p.148), not losing sight of discrete-point component from a traditional skills-based perspective.

Another insightful research in this line for developing future rating scales is a study done by Iwashita, McNamara, and O'Hegan (2008). They investigated 40 speech samples of IBT TOEFL to distinguish features determining performances by level. The criteria used for analysis were mainly divided into three categories: linguistic resources, phonology, and fluency. But the features included in each category were comprehensive: 1) grammatical accuracy, grammatical complexity, vocabulary in linguistic resources, 2) pronunciation, intonation, and rhythm in phonology, and 3) fill and unfilled pauses, repair, pausing time, speech rate, and mean length in fluency. Interestingly, features from all three categories affected the overall score. There was no single category more salient than others. The distinguishing features, however, were grammatical accuracy, vocabulary, pronunciation, and fluency.

Fulcher (1996) focused on the speaking construct of fluency by investigating whether the feature of fluency could discriminate well between students. 21 interview samples of English learners were examined. Results displayed 95% of the correspondences between the frequency counts of fluency and assigned students' band level.

Noticeable throughout the aforementioned investigations is that the common feature observed as most salient was fluency. It is assumed that however, discrepancies reveled in the results of the distinguishing criteria in the previous research are due in part to the idiosyncratic nature of the context where the test and rating were administered. This study attempts to cover what features of criteria were more influential in resulting in the overall level of performance and whether the speech competition context of this study had an effect on the rating, compared with those of aforementioned studies.

III. RESEARCH METHODS

1. Participants

A total of 35 Korean elementary and middle school students participated in an English

Speech Competition held by a university in Asan, Chungnam, and 11 native speaker lecturers working at a university in Chungnam rated their speaking performances. The students participating in the competition were composed of nine middle school students, thirteen 5 and 6th graders, and eleven 3rd and 4th graders. Of the 35 students participated, only 22 students' speeches were adopted for this study, excluding the speeches of thirteen third and fourth graders. It was announced to the applicants that students with the following conditions were not allowed to enter the competition: having experience living in English speaking countries more than 6 months, students awarded prizes in the previous competitions held by the same institution Table 1 provides a summary of student variables.

TABLE 1

Summary of Student Variables

	N	Average TOSEL score(L/S score*)	Gender F	Gender M
5-6th graders (elementary)	13	94(49)	6	7
Middle school students	9	77.2(39)	4	5

* L/S refers to listening and speaking section.

The native raters participating in the competition were all lecturers teaching English Conversation and other English courses such as presentation and extensive reading at university. They all varied in age, major, and the years of teaching experience as shown in Table 2. As this competition has been held annually since 2009, ten lecturers had experience working as raters in this competition more than twice except the one new lecturer. In addition, as the same rating criteria have been used since the first competition, most of them were familiar with the criteria given and what was expected of them as raters.

TABLE 2

Summary of Rater Variables

	Category	N
Major	English	2
	Non-English	9
Gender	Male	10
	Female	1
Years of Teaching English	1-5 years	1
	6-10 years	5
	Over 10years	5

This study does not focus on the rater effect between native and non-native raters in assessing non-native oral performance. However, there appeared a tendency that non-

native raters exerted more strictness on grammatical accuracy while native raters placed more emphasis on content and fluency, with no apparent difference in other areas of rating between the two rater groups (Brown, 1995; Kang & Ahn, 2012; Lee, 2010; Zang & Elder, 2011). Results focusing on discrepancies between the two rater groups, in general, report harsher ratings among non-native raters compared to those of their counterparts (Fayer & Krasinsky, 1987; Kang & Lee, 2012).

2. Data Collection

The data used for this study consist of 21 students' speeches, ratings of 11 native speaker lecturers on the students' performances in their speeches and questionnaires regarding the rater's perceptions of rating criteria.

1) Procedures

In order to enter the competition, students were required to apply online approximately one month prior to the competition. Upon applying, they were informed of the procedures they had to follow on the application online site. They went through the first screening 10 days after their application, by taking a Test of the Skills in the English Language (TOSEL). TOSEL is a proficiency test consisting of 2 sections: 1) listening and speaking, and 2) reading and writing. A total of 117 students applied and only 26 students (approximately 22%) were selected through the screening. Among the 26 selected students, 5 students did not turn up on the final competition. The cut score for the screening was 90 (elementary 5-6th graders), and 65 (middle school students) out of 100 as a perfect score. Table 3 illustrates what had been done in the preparation period for English Speech Competition.

TABLE 3

Preparation Process of English Speech Competition

Date	Work to do
Oct. 1-14	Application for English Speech Competition (online)
Oct. 25	Administering a screening test (TOSEL)
Oct. 31	Completion of accepting submissions of 'speech scripts' from students
Nov. 1-Nov. 10	Developing questions based on students' scripts
Nov. 1-Nov. 10	Developing questions from textbooks (Picture description/Wh-questions)
Nov. 21	Rehearsal
Nov. 22	Holding an English Speech Competition

The screened applicants were given a topic for their speech. The topic for this year was 'Travel'. Students were then asked to submit their scripts on the topic. After

accepting their submission of scripts, two lecturers were assigned a duty of making 2-3 questions based on the scripts students submitted. Other questions from the students' textbooks were also developed by two lecturers; one type of question was picture description for elementary school students and the other type was a wh-question (i.e. What healthy habits do you have?) for middle school students. The underlying reasoning behind preparing these questions was first to measure students' speaking ability for real communication. Therefore, based on the questions developed, students were asked to perform two more tasks apart from speech: namely, Interview for Speech, and Picture Description (elementary school students only)/Textbook Questions and Answers (middle school students only). Details on the tasks are summarized in Table 4.

TABLE 4

3 Tasks Given in the English Speech Competition

Type	Task details
1.Speech	● Students make a speech based on the script made. 2 minutes (middle school students)/1 min. 30 sec. (5-6th graders)/ 1 min. (3-4th graders)
2.Interview for speech	● After listening to students' speech, a questioner asks 2-3 questions prepared based on their scripts and students are asked to answer the questions. (2 min. for everyone)
3-1.Picture Description (Elementary only)	● Students are given a picture taken from their textbooks and asked to describe the picture. (1 minute)
3-2.Textbook Question and Answer (Middle only)	● Students are asked to answer the questions taken from their textbooks. (1 minute)

Rating was done in real time for the 3 tasks; 8 raters participated in the rating of middle school students' speech performances, and 4 raters did for elementary school students. Students made their speeches for a given time (2 minutes for middle school students/ 1 min. and a half for elementary school students), standing at the front of a lecture hall.

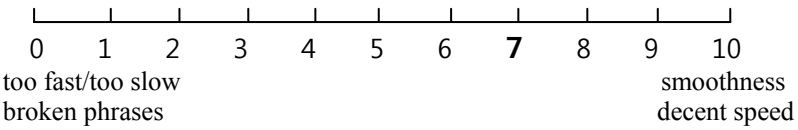
2) Instrument for Rating Criteria

The rating scale used for this study is composed of six criteria: *fluency, grammar, pronunciation, presentation, content,* and *memorization/timing* (Appendix A). The criteria was developed adopting primarily the list made by Brown (2007). Also, one more criterion, *memorization/timing* was added, considering the competition setting peculiar to this study. The criteria used for each task differed as the expected behaviors for each task were not the same. Details of the criteria for each task will not be included here as this study deals solely with the results of assessment on the speech task for analysis.

Some of the criteria were weighted; raters were asked to give a mark on the 10-point Likert scale on the criteria: *fluency, grammar, pronunciation* and *presentation*, whereas 5-point Likert scale was given to the other two criteria, *content* and *memorization/timing*. In total, students could obtain 50 at maximum for the speech task. As indicated in the participants section, most of the raters except one new lecturer were aware of what was expected of them to rate speech as they had experienced rating for the competition using the same criteria. Nevertheless, a number of questions were raised on a rehearsal from the lecturers as for how to use the criteria. They made an agreement on certain features on *fluency* and *grammar*. However, some of the questions were left unsolved. For example, there were confusions among the raters regarding *bad pronunciation vs*. *good pronunciation*. There was an opinion that less weight should be assigned to *pronunciation*, without explicit and detailed descriptions of the criterion. Instead, it was suggested more weight be assigned to *content*. However, it was agreed to reflect the issues raised on the rating criteria for the next competition due to the tight schedule. The following shows examples of simple explanations given on the rating sheet for each criterion:

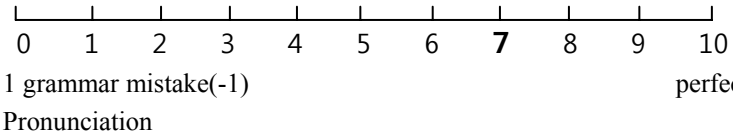
1. Speech

a) Fluency



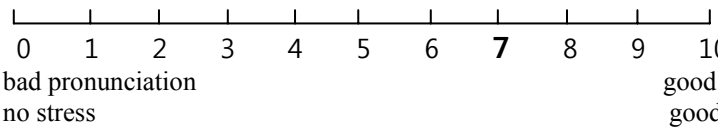
too fast/too slow — smoothness
broken phrases — decent speed

b) Grammar



1 grammar mistake(-1) — perfect grammar

c) Pronunciation



bad pronunciation — good pronunciation
no stress — good stress
no intonation — good intonation

3) Questionnaire

Following the rating, questionnaires were completed by the raters. Out of 11, only 9 questionnaires were returned and reflected for analysis of this study. Questions included were as follows; 1) which of the following criteria do you find important in assessing

speaking skills? 2) which of the following features do you think crucial for mutual understanding when a non-native English speakers talks? For the first question, they marked their score on a 5-point Likert scale for the degree of importance they assigned, while for the second question, they were asked to choose five most important features and number them in the order of importance (Appendix B). The questions were developed to see how their perceptions of the importance of each criterion were reflected in their rating of the criteria. Also, the latter question was included for two reasons. Questions on rating criteria on rehearsal were mostly centered on pronunciation, and the scores they assigned on pronunciation appeared inconsistent compared with other types of criteria. The second question was, therefore, designed to find out how the raters as lecturers as well perceived relative importance of each pronunciation feature given.

The pronunciation features chosen for the second question were adopted from a list summarized by Jenkins (2002). Jenkins (2002) formed a pronunciation core for teaching English as a Lingua Franca as a guide for Lingua Franca interactions, not interactions between native and non-native speaker of English. Its features include: 1) all the consonants, 2) consonant clusters, 3) the contrast between long and short vowels, and 4) nuclear stress- a stress on the most important words in a group of words. These four features were all provided in the questionnaire as well as other six features regularly taught on English pronunciation courses:1) the '*th*' sound, 2) vowel quality, 3) weak forms such as schwa, 4) features of connected speech such as assimilation, 5) word stress, and 6) pitch movement. According to Jenkins (2002; 2011), the latter six features are considered all important for a native speaker listener, but not as essential as the former features formed as a Lingua Franca core for intelligibility in interactions between non-native speakers of English.

3. Data Analysis

The data analysis for this study used correlation statistics to answer the first research question on what criteria are most influential in determining final scores for speech performances. The scores obtained from each criterion were compared to those of other criteria and the final scores using correlation statistics. Then they were correlated with TOSEL scores students received on the screening. TOSEL is a proficiency test widely taken especially by relatively young learners of English. As this study focused on speaking ability, only the scores of the listening and speaking section were compared with the data of this study. On the TOSEL, question types for speaking consist of multiple-choice types: test-takers are asked to choose the best answer for a given question, a dialog and a picture-cued description. A concern may be raised in correlating scores of speaking performances and those of listening and speaking skills as the

speaking test on the TOSEL is not based on actual oral performance. However, if the scores between the two data sources are correlated, then it may be cautiously suggested that the validity of the used rating criteria for not only speaking but only listening is approved. For the analysis of the questionnaire results, descriptive statistics were used for mean score and frequency counts.

IV. RESULTS

1. Rating Criteria Determining Speaking Performance

Descriptive statistics of the rating scores of each criterion and final scores are presented in Table 5, indicating that the data of this study were normally distributed overall. On the whole, the scores of middle school students were lower than those of 5th and 6th graders in each of the criterion. The highest average score was given to *presentation* in the two data, while the lowest average score was given to *grammar* in the speeches of middle school students, *pronunciation* for 5th and 6th graders. It may be assumed that the context effect was reflected in the results of higher scores given to *presentation,* the features of which include non-verbal elements of interactions such as making eye contact, being confident and loud voice projection. It is possible that the students preparing for the speech competition were made more aware of the significance of such elements of *presentation* of giving speeches in front of audiences than those who perform other kinds of speaking tasks.

TABLE 5

Descriptive Statistics of the Results of Rating

	5th-6th Graders (N=13)			Middle School Students(N=9)		
	Mean	SD	Skewness	Mean	SD	Skewness
Pre	8.19	.87	-.15	7.25	1.4	.79
Con	4.34	.85	.61	3.36	.91	.31
Mem	4.30	.59	-.14	3.47	1.47	-.91
Flu	6.96	1.05	-.96	5.45	2.21	-.97
Gram	7.26	.88	-.98	4.75	1.04	.23
Pron	6.57	1.03	1.13	5.9	1.08	-.67
Scores	37.65	3.41	-.62	30.2	6.79	.10

*Presentation (Pre), Content (Con), Memorization (Mem), Fluency (Flu), Grammar (Gram), Pronunciation (Pron)

As reported in Table 6, correlation matrix in the data of 5th and 6th graders shows each of the six criteria, except *pronunciation*, was significantly correlated with the final scores, with *r* value ranging from 0.604 to 0.901. Specifically, *fluency, content* and *grammar*

were of large effect sizes in correlations with the final scores, with *fluency* most highly correlated with the final scores. Moreover, a moderate correlation was found between the final scores and the TOSEL scores (0.597), while *content* solely showed a correlation with the final scores among other criteria (0.657). There were moderate and strong correlations among the four criteria, *presentation, content, memorization/timing,* and *fluency,* whereas *grammar* and *pronunciation* were not related with each other and other criteria. What is noticeable, however, is a relatively strong correlation found between *grammar* and *fluency* with *r* value 0.756.

TABLE 6

Correlation Matrix by Criteria and TOSEL(5-6 Graders)

	Pre	Con	Mem	Flu	Gram	Pron	Scores	TOSEL
Pre	1	.809**	.647*	.641*	.358	-.159	.799**	.536
Con		1	.597*	.597*	.552	-.112	.845**	.657*
Mem			1	.609*	.303	-.656*	.604*	.292
Flu				1	.756**	.124	.901**	.374
Gram					1	.398	.815**	.468
Pron						1	.148	.105
Scores							1	.597*
TOSEL								1

*p<.01 **p<.05

Table 7 shows the degree of correlations among the three areas: between the criteria, between the criteria and final scores, and between the criteria/ final scores and TOSEL scores from the speech performances of middle school students. Similar to the results gained from the speech performances by 5th and 6th graders, data of middle school students displayed conspicuous correlations between the criteria except for *pronunciation* and the final scores with *r* value from 0.743 to 0. 944. In particular, highest correlations were observed between *fluency* (0. 944), *presentation* (0.924), *content* (0.923) and the final scores.

TABLE 7

Correlation Matrix by Criteria and TOSEL(Middle School Students)

	Pre	Con	Mem	Flu	Gram	Pron	Scores	TOSEL
Pre	1	.892**	.735*	.812**	.623	-.527	.924**	.608
Con		1	.743*	.789*	.783*	.345	.923**	.633
Mem			1	.946*	.385	.089	.890**	.751*
Flu				1	.526	.253	.944**	.891**
Gram					1	.554	.703*	.583
Pron						1	.433	.351
Scores							1	.807**
TOSEL								1

* p<.01 **p<.05

With respect to correlations with TOSEL scores, TOSEL scores of the listening and speaking section presented a strong correlation with the total scores (0.807), and with two criteria, *fluency* and *memorization/timing*. There also appeared significant relationships between the criteria: *presentation, content, memorization/timing,* and *fluency*, with the highest correlation between *fluency* and *memorization/timing* (0.946). No relationships were found in the results of two criteria, *grammar* and *pronunciation* between the two and with other types of criteria, with one exception of a moderate correlation between *grammar* and *content* (0.783).

2. Raters' Perceptions of the Rating Criteria

1) Which of the following criteria do you find important in assessing speaking skills?

In the questionnaire carried out after the competition, 9 raters responded to a question asking about the degree of importance they assign to each of the rating criteria of speaking assessment. The criteria perceived very important (5.0 points) to important (4.0 points) by the raters were discovered as *fluency* (4.4), *content* (4.3), *grammar* (4.2), and *pronunciation* (4.0), while *presentation* (3.2), and *memorization* were (3.1) considered not as important as other types of criteria.

2) Which of the following features do you think crucial for mutual understanding when a non-native English speakers talks?

The raters were provided with ten pronunciation features to choose five most important features from. The numbers on the vertical line of Table 8 indicate the features included: 1. all the consonants, 2. the contrast between long and short vowels, and 3. the '*th*' sound, 4. vowel quality, 5. consonant clusters at the beginning or in the middle of the word, 6. nuclear stress- a stress on the most important words in a group of words, 7. weak forms such as schwa, 8. features of connected speech such as assimilation, 9. word stress, and 10. pitch movement.

Table 8 presents the results of the questionnaire. Inconsistent responses were observed throughout the questionnaire data among the raters. For example, one rater chose 'the '*th*' sound' as the most important one while seven other raters did not even choose the feature as important at all. Through personal communication with some of the raters, they said it was difficult to choose features more important as all the features were considered important in similar degrees. However, Feature 6 (nuclear stress-25 points),

Feature 1 (all the conconants-21 points), Feature 4 (vowel quality-19 points) obtained highest points, followed by Feature 9 (word stress-18 points), and Feature 2 (short vs. long vowel-14 points).

TABLE 8

Scores Assigned to Pronunciation Features by Rater

Feature	R1	R2	R3	R4	R5	R6	R7	R8	R9	**Score**
1	*0*	*5*	*5*	*0*	*0*	*0*	*4*	*2*	*5*	***21***
2	*0*	*4*	*4*	*5*	*2*	*1*	*0*	*0*	*0*	***14***
3	*0*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*1*	***2***
4	*3*	*3*	*0*	*4*	*5*		*3*	*1*	*0*	***19***
5	*0*	*0*	*3*	*0*	*1*	*2*	*0*	*0*	*0*	***6***
6	*5*	*2*	*0*	*0*	*3*	*4*	*5*	*5*	*1*	***25***
7	*2*	*0*	*0*	*0*	*0*	*5*	*2*	*0*	*3*	***12***
8	*1*	*4*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	***5***
9	*4*	*1*	*2*	*1*	*0*	*3*	*0*	*4*	*4*	***18***
10	*0*	*0*	*1*	*3*	*4*	*0*	*1*	*3*	*0*	***12***

*Rater(R)

*Numbers(italics) in the table refer to points given to the each feature by the order of importance (i.e. 5 points were given for the important feature marked as 1 on the questionnaire, 4 points for the second important feature 2…).

V. DISCUSSION

1) How does each of the rating criteria relate to each other to contribute to the final score of students' performance?

The first research question deals with the ways of each criterion works to contribute to the students' overall performance. The correlation analyses reported that all the criteria except *pronunciation* showed remarkable relationships with the final scores of elementary and middle school students, suggesting that *presentation, content, memorization, fluency,* and *grammar* acted as valid predictors of students' overall speaking ability. However, in predicting the final score, most salient criteria common in both data sources were *fluency* and *content*, with a moderate correlation between each other. The findings correspond to those of studies done by Tajeddin et al. (2011) and Fulcher (1996), despite the difference in the task type used for each study. In particular, consistent with the results of Fulcher (1996)'s study, the findings of this study may offer a possibility that even a single criterion or a couple of criteria can successfully predict learners' final scores, which may lead raters to relieve their cognitive overload by rating more than 7 criteria. The five criteria displaying considerable to moderate relationships with the final score were also correlated with one other with varying degrees of

correlations, but *grammar* differed in the way of relating with other criteria.

In the speech of elementary school students, *grammar*, apart from the two strong criteria, played an important role in resulting in the overall performance, compared with its relatively moderate role in the speech of middle school students. *Grammar* was correlated solely with *fluency* in elementary school speeches and *content* in the speeches of middle school students. It may be suggested that *grammar* probably made the relatively significant contribution to the 5th and 6th graders' speeches by relating with the most powerful criterion, *fluency*. Elementary school students are considered to have less difficulty in processing sentences as they had less complex and shorter sentences in their speeches than middle school students, in addition to the shorter time given to them for speeches. As a result, it is expected few grammatical mistakes were made while making speeches if they had practiced hard enough to be fluent. Yet, in the case of middle school students, there appears to be higher possibilities that they make more grammatical mistakes if they were processing complicated sentences while speaking, with the effect of practice being less obvious. Indeed, of the criteria, the lowest score in average was given to *grammar* of the middle school students' speeches. Or with respect to its observed relationship with *content*, there is also a possibility that *grammar* can suffer if *content* itself was not well organized.

Another noticeable finding of this study was detected in the rating of *pronunciation*. *Pronunciation* was not correlated with the final score and with any of the criterion, not supporting results of the previous research conducted by Plough et al. (2010) and Iwashita et al. (2008). Possible reasons for this will be discussed in detail in the following section dealing with raters' perceptions of the criteria.

2) Is there any correlation between the criteria and the TOSEL score of the listening and speaking section? If any, how do each criterion, or relations between the criteria, contribute to the TOSEL score?

There were correlations between the final score of speeches and the TOSEL score reflecting the overall listening and speaking ability in English. The link emerged in this study may provide an evidence to validate the criteria used and the way of using the criteria for assessing speaking ability. However, a caution could be made in interpreting this link as a way to prove the validity of the criteria in terms of the relevance and use of each criterion for speaking assessment. While the TOSEL score has been widely used in Korea as an official score to represent the overall proficiency in the four skills of English, the speaking part is assessed in an indirect way, not orally (i.e. choosing the right response to the previous question provided by listening materials). According to Hughes (2003) indirect assessment may lead to more precise evaluation of a skill to be assessed

by testing an indirect item representative of a certain skill. However, he goes on to suggest that there is a tendency that the connection between indirect test item and a skill is rather weak and uncertain.

Regarding the relative contribution of each criterion to the TOSEL score, *fluency* and *memorization* displayed correlations in the speech performances of middle school students, but *content* only in the speeches of 5th and 6th graders. Inconsistent results between the two groups cannot be explained here fully, but it is only assumed that in middle school students' speeches, *fluency* as the most powerful predictor of a final score acted similarly for the TOSEL score by relating with *memorization* revealing the most robust correlation *fluency* (0. 946) among others. Indeed, a more significant relationship was found in the speeches of middle school students.

3) How do native speaker teachers perceive the relative importance of each rating criterion?

Questions provided in the questionnaire were twofold: 1) their perceptions of the relative importance of each criterion for speaking assessment and 2) their perceptions of distinguishing pronunciation features. The criteria with high scores on the marking of importance were *fluency, content, grammar*, and *pronunciation*. This finding may suggest that non-native speaker raters' perceptions were reflected in their rating of the criteria on the whole, noting that *fluency, content,* and *grammar* were found as influential indicators representing overall speaking ability of the students in this study.

Meanwhile, as reported, *pronunciation* did not work as a useful criterion in determining students' overall speaking ability. In other words, pronunciation was perceived as relevant for assessing speaking but turned out to be neither relevant nor reliable in this study. *Pronunciation* was not correlated with any of the criteria and the final score. The reason for this could be attributed to the differing ways of measuring pronunciation among the raters. The findings of the questionnaire report that raters had all different ideas on the relative importance of the ten pronunciation features and it is possible that these differences were reflected in the process of measuring the students' *pronunciation* on the competition, leading to no correlations with any of the criteria. The pronunciation features perceived important were nuclear stress (Feature 6), all the consonants (Feature 1), vowel quality (Feature 4), word stress (Feature 9) and short vs. long vowel (Feature 2) in the order of importance. Yet, of the five features perceived important, three features (Feature 1, 2 and 6) overlapped with a Lingua Franca pronunciation core suggested by Jenkins (2002, 2011), implying that the raters might have a certain degree of understanding on the notion of Global English.

For *pronunciation* to be more relevant in assessing speaking skills, a more detailed

explanation on how it should be measured needs to be made based on the current demands or practices in the teaching of pronunciation and agreement by the raters. As O'Sullivan (2008) points out, when designing rating criteria, contexts where an assessment tool is to be used should be taken into consideration. Learners of English in EFL contexts probably have higher possibilities that in the future they interact more with non-native speakers of English than native speakers. Then when it comes to pronunciation, 'intelligibility', instead of a certain country's native speaker pronunciation model becomes an issue (Jenkins, 2002; Knight, 1992; McNamara, 2000). It is thus recommended that raters, regardless of whether they are native or non-native speakers, should be made alert of the need of such features as Lingua Franca core for intelligibility so that those features are more effectively reflected when rating non-native students' speaking performances.

VI. CONCLUSION

This study has attempted to find out how effective the employed criteria for this study were in deriving the final score of speaking performances by Korean secondary school students in a speech competition. Also, the raters' perceptions of the criteria were investigated to see how native raters' perceptions of the criteria were reflected in their rating. First and foremost, all the criteria except *pronunciation* displayed significant correlations with the final scores of the students' speeches, validating the relevance of the criteria used in this study. Specifically, the most salient criteria attended to by the raters were indicated as *fluency* and *content*. Second, it was found that on the whole, the raters' perceptions of the important criteria corresponded with the results of their actual rating of the criteria. *Fluency, content, grammar* and *pronunciation* were perceived as essential ones in contributing to students' overall speaking performances. Third, there emerged a relationship between the final score and the listening and speaking score of TOSEL, improving the degree of validity of the criteria. Fourth, *pronunciation* did not work as a useful criterion in affecting students' overall speaking ability by not correlating with any of the criteria and final scores. A possible reason for this may be attributed to inexplicit explanations provided for rating *pronunciation,* along with the raters' inconsistent perceptions of the relative importance of pronunciation features. A need for more explicit guidelines with examples is strongly recommended.

Even though the current study tried to make practical implications using the empirical data of students' speeches, there are limitations of this study. To point out a few, first, the sample size used in this data is small, so the findings of this study should be applied to other populations with some caution. Further studies on the assessment of speaking

using rating criteria should be conducted with higher sample sizes in EFL to corroborate implications made in this study. In addition, to suggest the validity of rating criteria more effectively, it is recommended to adopt a more direct assessment (i.e. holistic assessment) tool to compare the final scores of this study with.

REFERENCES

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12,* 1-15.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*: White Plains, NY: Pearson Education.

Brown, H. D. (2007). *Teaching by principles: An Interactive approach to language Pedagogy*. White Plains, NY: Pearson Education.

Eckes, T. (2008). Rater types in writing performance assessment: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Fayer, J.M. and Krasinski, E. (1987). Native and nonnative judgements of intelligibility and irritation. Language Learning 37.3, 313-326.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208-238.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Iwashita, N. Brown, A. McNamara, T. & O' Hagen. (2008). Assessed levels of second language proficiency. How distinct? *Applied Linguistics, 29*, 24-49.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an International Language. *Applied Linguistics 23*, 83–103.

Jenkins, J. (2011). *Global English and the teaching of pronunciation*. Teaching English:, British Council. Retrieved November 10, 2014, from the World Wide Web: http://teachingenglish.org.uk/article/global-english-teaching-pronunciation.

Jin, T. & Mak, B. (2012). Distinguishing features in scoring L2 Chinese speaking performance: how do they work? *Language Testing, 31*(1). 23-47.

Kang, S.-H. & Rhee S.-C. (2011). A study on the suprasegmental parameters exerting an effect on the judgment of goodness or badness on Korean spoken English. *Malsori and Speech Science 3*(2*)*, 3-10.

Kang, S. & Ahn, H. (2012). A comparative study on criteria and tasks in Korean English speaking assessment by native and non-native raters. *Language Research, 48*(2). 241-262.

Kim, H. Y (2003). The types of speaking assessment tasks used by Korean junior secondary school English teachers. *Asian EFL Journal. December,* 1-30.

Korea Institute of Curriculum and Evaluation. (2005). *Youngeogoa Kyouykkoajeoung Kaesonbanganeongoo. Korea*: The Ministry of Education and Science.

Knight, B. (1992) Assessing speaking skills: a workshop for teacher development. *ELT Journal. 46*(3), 294-302.

Lee, C.-H. (2010). Improving inter-rater reliability in oral proficiency test at college level. *Modern Studies in English Language & Literature 54*(1). 367-387.

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

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. (2000). *Language Testing*. New York, Oxford University Press.

O'Sullivan, B. (2008). Notes on assessing speaking. Retrieved November 10, 2014, from the World Wide Web: http://www.lrc. cornell.edu/events/past/2008-2009/osull1.pdf.

Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing, 25*(1), 127-143.

Plough, I. C., Briggs. S, L. & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27*(2), 235-260.

Tajeddin, Z., Alemi, M, & Pashmforoosh, R. (2011). Non-native teachers' rating criteria for L2 speaking: Does a rater training program make a difference? *TELL, 5*(1), 125-153.

Zhang, Y. & Elder, C. (2011). Judgements of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing 28*(1), 31-50.

APPENDIX A

Rating Sheet

Middle 1-3 Graders	
Contestant Number	

Judge Name _____

1. Speech 1

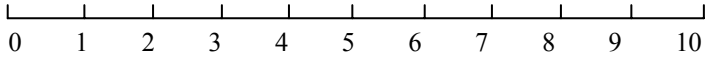
a) Fluency

0 1 2 3 4 5 6 **7** 8 9 10

too fast/too slow — smoothness

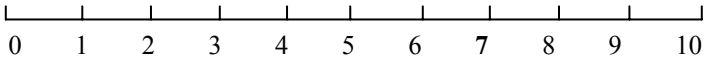
broken phrases — decent speed

b) Grammar



1 grammar mistake(-1) perfect grammar

c) Pronunciation

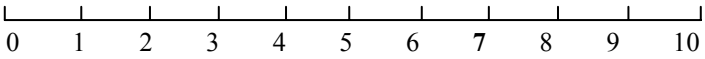


bad pronunciation good pronunciation
no stress good stress
no intonation good intonation

Subtotal:
/30

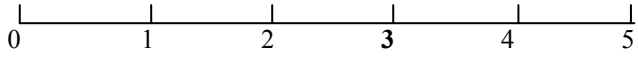
2. Speech 2

a) Presentation/Stage Manners



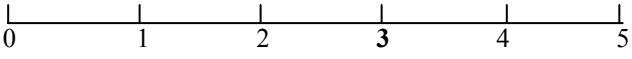
No eye-contact, shy good props (+1), confident
Fumbling of props good voice loudness
No smile good eye-contact, pleasant

b) Content



not interesting interesting
not well-organized well-organized

C) Memorization/Timing



hesitation
prompting (word(-1),
phrases (-2), sentences (-3))

Subtotal:
/20

APENDIX B
Questionnaire

Name _____
Nationality _____
The years of teaching English in Korea _____
The years of teaching English in ESL/EFL contexts _____

1) presentation (eg. eye-contact)	1	2	3	4	5
2) content	1	2	3	4	5
3) memorization	1	2	3	4	5
4) fluency	1	2	3	4	5
5) grammar	1	2	3	4	5
6) pronunciation	1	2	3	4	5

2. Which of the following pronunciation features do you think are crucial for mutual understanding when a non-native speaker of English talks? Choose the 5 most important features and number them in the order of importance (1 –most important …… 5 - least important).

1) all the consonants _____
2) the contrast between long and short vowel (eg 'sit' vs. 'seat') _____
3) the 'th' sound (eg 'thin' or 'this') _____
4) vowel quality where the length is not involved (eg 'bad') _____
5) consonant clusters at the beginning or in the middle of the word _____
   (eg 'string' cannot be simplified to 'sting' or 'tring')
6) the stress on the most important word in a group of words _____
   (eg 'My SON uses a computer.')
7) schwa instead of their full quality (eg 'to' and 'from) _____
8) assimilation (eg 'red paint' becomes 'reb paint') _____
9) word stress _____
10) pitch movement _____

Thank you very much for your cooperation ^^**********

Examples in: English

Applicable Languages: English

Applicable Levels: Secondary and college

Hera Chu
Dept. of Foreign Language Education
Sun Moon University
70 Sunmoon-ro 221
Tangjeong-myeon, Asan
Chungnam. 336-708. Korea
Tel: 82-41-530-2543
Email: hera_chu@yahoo.co.kr

Received in January 15th, 2015
Reviewed in February 26th, 2015
Revised version received in March 5th, 2015