TELL, Vol. 5, No. 1, 2011, pp. 125-153

# Non-native teachers' rating criteria for L2 speaking: Does a rater training program make a difference?

**Zia Tajeddin**[1]
*Associate Professor, Allameh Tabataba'i University*
**Minoo Alemi**
*Ph.D. in TEFL, Sharif University of Technology*
**Roya Pashmforoosh**
*M.A. in TEFL, Sharif University of Technology*

*Received August 18, 2011*
*Accepted October15, 2011*

## Abstract

Inconsistent rating seems to emanate largely from the application of different rating criteria. It follows that rater training programs may bring about higher rating consistency. This study aimed to explore non-native EFL teachers' rating criteria for L2 learners' speaking performance and to measure the impact of a rater training program on raters' rating criteria. As many as 28 EFL teachers rated 10 monologs both before and subsequent to a rater training program and specified the criteria they applied in their ratings. The findings show they specified 10 common rating criteria, ranging from fluency to communicative effectiveness. However, they reconsidered the significance of a few criteria after the program. While there was a sharp decline in the significance given to the rate of speech and affective variables, the training program led to

[1] Corresponding author at: Allameh Tabataba'i University
E-mail Address: zia_tajeddin@yahoo.com

the rising importance of fluency, comprehension, and organization. The results reveal that the traditional skills-and-components-based perspective on language proficiency makes teachers lose sight of macro-level, higher-order components like fluency and organization. To conclude, the effective rating of language skills needs to be embedded in teacher education programs.

**Keywords:** English teacher education, assessment, speaking, rating criteria

## 1. Introduction

Rating criteria play a significant role in rater-mediated speaking assessment. This is true of various rating scales, i.e. analytic vs. holistic, that are applied to the assessment of L2 speaking. Given the prominent status of rating scales, several studies have examined the differences in the raters' perception and use of speaking rating criteria (e.g. Barnwell, 1989; Hadden, 1991; Kim, 2009; Plough, Briggs, & Van Bonn, 2010; Zhang & Elder, 2011). Rater variability as a result of differential rating perceptions is becoming increasingly important because different raters may draw on their own personalized constructs irrespective of the criteria they are given, and hence "it would be a mistake to assume that high inter-rater reliability constitutes evidence of the construct validity of the scales or performance descriptors that are used" (Brindley, 1991, p. 157).

Gaining a better understanding of the speaking construct requires empirical research to explore teachers' criteria when assessing L2 learners' speaking ability and to discover the impact of a training program focused on rating speaking on raters' rating criteria. A rater training program on rating criteria for L2 speaking is among those teacher education courses that contribute greatly to a development in teachers' rating ability. The question that still remains underexplored in the literature on speaking assessment is what criteria teachers use for rating speaking and what can be done to increase consistency in raters' rating criteria. As assessing

speaking involves human raters to judge and score examinees' performances (Brown, 2004) and speaking is an important language skill to assess (Bachman, Lynch, & Mason, 1995; Brown, 1995; Chuang, 2009; Kim, 2005; Lumley & McNamara, 1995; Luoma, 2004; Underhill, 1987), the criteria EFL teachers use for rating speaking and the significance of each were the main focus of the present study.

## 2. Literature Review

In this section, the nature of speaking assessment, research on rater variables in rating speaking performance, and rater training for speaking assessment will be reviewed.

### 2.1 Speaking Assessment

The assessment of speaking is a complicated matter due to a variety of factors that may affect final rating. Understanding the nature of speaking not only helps define the construct in question but ultimately makes it possible to identify the factors involved in speaking assessment (Kim, 2010). To define speaking ability, the componential nature of L2 oral ability has been examined to determine the features distinguishing performances at each level (Chalhoub-Deville, 1995; Iwashita, Brown, McNamara, & O'Hagan, 2008). A number of researchers analyzed such features as linguistic resources (grammatical accuracy and vocabulary), phonology (pronunciation and intonation), fluency (speech rate and length of utterance), and content (ideas and completeness of response) of L2 speaking (e.g. Chuang, 2009; Plough et al., 2010; Zhang & Elder, 2011).

Over the past two decades, L2 speaking assessment has been directed toward the use of more oral performance tests which require learners' actual performance in a simulated language use context. Such tests typically involve three variables: examinees' performance, task specification, and raters' judgment. In terms of the schematic representation of interaction in performance assessment of the speaking skill which was primarily presented by

Kenyon (1992) and later on developed by McNamara (1995), there is an interaction between candidates, tasks, and interlocutors in the performance phase, and between raters and rating scales in the assessment phase. Regarding the interaction between raters and rating criteria, rater variability can manifest itself in different ways. According to Eckes (2008, p. 156), raters may vary in:

(1) the degree to which they comply with the scoring rubric,
(2) the way they interpret criteria employed in operational scoring sessions,
(3) the degree of severity or leniency exhibited when scoring examinee performance,
(4) the understanding and use of rating scale categories, or
(5) the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks.

Just as important and critical as the selection of criteria used in oral performance assessment is the choice of other factors that affect such assessment, including candidate (Lumley & O'Sullivan, 2005; Nakatsuhara, 2011; O'Loughlin, 2002), task (Chalhoub-Deville, 1995; Elder, Iwashita, and McNamara, 2002; Fulcher & Marquez Reiter, 2003; Norris, Brown, Hudson, & Bonk, 2002; Robinson, 2001; Shohamy, 1994; Wigglesworth, 1997), interlocutor (Brown, 2003; O'Sullivan, 2002), and rater (Brown, Iwashita, & McNamara, 2005; Barnwell, 1989; Brown, 1995; Eckes, 2005; Elder, 1993; Kim, 2009; Lumley, 1998; Lumley & McNamara, 1995; Lynch & McNamara, 1998; Wigglesworth, 1993). Therefore, it is important to understand the relative contributions of these factors, including the rater variables, to the final scores assigned to speaking performance.

## 2.2 Research on Rater Variables in Rating Speaking Performance

Inconsistency in raters' rating criteria for assessing L2 speaking and hence rater variability is one of the main themes in research on rating speaking. Diverse rater groups may apply a different set of criteria with which they assess L2 learners' speaking ability

(Chalhoub-Deville, 1995; Shohamy, Gordon, & Kraemer, 1992; Zhang & Elder, 2011). The results of studies (e.g. Barnwell, 1989; Galloway, 1980; Hadden, 1991) indicate that teacher raters were more critical of linguistics aspects of learners' speaking ability in comparison with non-teaching raters. In an influential study, Galloway (1980) made a comparison between two groups of non-teaching native speakers who were the residents of their own community or the learners' community. It appeared that non-teaching native speakers residing in the learners' community were more tolerant of L2 learners' performances. In fact, they differed from those residing in the target language community in terms of their applied speaking rating criteria. Regarding the native and non-native differences in judgment, Fayer and Krasinski's (1987) study was an attempt to deal with non-native speakers who were also found to be less tolerant of errors in comparison with their native counterparts. To arrive at a better value judgment, as Chalhoub-Deville (1995) argues, an empirical investigation is indeed needed to derive the criteria salient to different rater groups in judging learners' L2 speaking ability.

Studies on speaking performance assessment have revealed that raters approach the task of evaluation with different levels of severity/leniency. For instance, Ang-Aw and Goh (2011) suggested that raters referred to a wide range of criterion factors (e.g., elaboration of response, clarity of expression, and engagement in conversation) and non-criterion factors (e.g., novelty of ideas, range of vocabulary, and inter-candidate comparison) when assessing learners' speaking performance. The tendency toward great variation among teacher raters has also been documented for native and non-native EFL teachers' evaluation of L2 learners' speaking performance (Barnwell, 1989; Brown, 1995; Kim, 2009; Zhang & Elder, 2011). Kim's (2009) study showed that native raters tended to focus more on elaborate features of speaking than those of non-native raters in the areas of pronunciation and specific grammar use. Similarly, the study carried out by Zhang and Elder (2011) revealed that notable differences between native and non-native teachers' judgments emerged with respect to the saliency assigned to the content features of oral communication like "relevancy to the topic" and "ideas." Native teachers appeared to be more concerned with

message-focused criteria, whereas non-native teachers were less concerned with specific instances of language use. Previous studies have also looked at untrained native vs. non-native teachers (Caban, 2003; Fayer & Krasinski, 1987; Kim, 2009; Zhang & Elder, 2011) and trained native vs. non-native teachers (Brown 1995). The findings suggested that untrained non-native teachers attended more to the linguistic resources as a justification for their scores and that native teachers drew more often on the non-linguistic categories of content and fluency. However, no consensus on the effect of rater training programs on teachers' rating criteria for L2 speaking has yet been reached.

## 2.3 Rater Training for Speaking Assessment

Several studies have been made to investigate disagreement between raters who may exhibit overall or particular patterns of harshness/leniency in relation to particular items and particular candidates (e.g. Bachman, Lynch, & Mason, 1995; Caban, 2003; Iwashita, McNamara, & Elder, 2001; Kim, 2009; Lumley & McNamara, 1995). There may be a kind of "rater-item" and "rater-candidate" interaction leading raters to over- or under-rate candidate performances (McNamara, 1996). Variation in the interpretation of rating categories and the use of a wide range of scales may lead raters to have divergent scores. In view of the central tendency, some raters just look for similarities between candidates to assign scores in the middle of the rating scales; however, others may just see the extreme differences between candidates that drive them to apply the end of the scales.

To establish common grounds on which raters come into agreement with one another, especially in terms of a common interpretation of rating scales and categories, rating programs are of significance. A number of studies have then been conducted to investigate the effectiveness of rater training programs in performance assessment settings (Elder, Barkhuizen, Knoch, & Randow, 2007; Kondo-Brown, 2002; Lumley & McNamara, 1995; MacIntyre, 1993; Shohamy et al., 1992; Weigle, 1994, 1998; Wigglesworth, 1993). They all emphasize two conclusions: (1) rater

training should be at the service of making raters more self-consistent; and (2) rater training should be implemented within the assessment context to eliminate rater variability up to a certain point; beyond the threshold level, elimination of differences is indeed neither desirable nor possible.

Rater-involved assessment engages subjective evaluations which need rater training programs to mitigate this subjectivity. It has been found that training serves to attenuate extreme differences between raters (Weigle, 1994, 1998; Wigglesworth, 1993; Xi & Mollaun, 2009). However, as Brown (1995) and Douglas (1997) argue, rater differences would still exist after training. Brown (1995) looked at scoring performance of trained native and non-native raters. It was found that the overall rater severity did not differ significantly after the rigorous training. It appears that raters may have different perceptions of a good performance and, at times, this could be due to raters having different interpretations of rating scale categories. Recent studies (Kim, 2009; Zhang & Elder, 2011) showed that raters from native and non-native language backgrounds assigned similar scores to the speaking samples for perhaps very different reasons. With regard to the effect of training on assessment procedures, the question is whether individual raters' agreement should be made perfectly complete or partially acceptable. With reference to the paradoxical condition under which reliability increases at the expense of validity, raters' rating criteria for L2 speaking require further exploration. Although EFL teachers are frequently involved in assessing speaking, there has been inconclusive evidence regarding the effect of training programs on their speaking rating criteria.

## 3. Purpose of the Study

To increase the quality of raters' performances in speaking assessment, a rater training program is needed to introduce teacher raters to rating guidelines. To investigate the effect of such a program, this study was conducted to discover the criteria non-native EFL teachers use to rate L2 speaking before and after a

speaking rater training program. The following questions constituted the main focus of the study:

(1) What criteria do non-native EFL teachers use to rate the speaking performance of EFL learners before and after speaking rater training?
(2) Is there any difference in the frequency and significance of the criteria non-native EFL teachers use to rate the speaking performance of EFL learners before and after speaking rater training?

# 4. Method

## 4.1 Participants

The participants of the study were 28 teacher raters and one rater trainer, as described below:

**Teacher raters:** All 28 non-native EFL teachers attending the rater training program were undergraduates and graduates of the English-related fields of study, including teaching English as a foreign language, English literature, and translation, as well as non-English majors of science and engineering who were teachers at two language institutes. They all voluntarily attended the speaking rating program which was designed to qualify them to score students' speaking ability. As displayed in Table 1, the teacher raters were different in terms of the course levels they taught at their institute, ranging from elementary to advanced courses. They were categorized into two groups of less experienced teachers (N=18), below 5 years of teaching, and more experienced (N=10), over five years. Table 1 depicts the relevant characteristics of the teachers.

**Table 1:** EFL teacher raters' profile summary

| Variables | Categories | Frequency | Percentage |
|---|---|---|---|
| Degree | BA | 16 | 57.1 |
| | MA | 12 | 42.9 |
| | Total | 28 | 100 |
| Major | English | 21 | 75.0 |
| | Non-English | 7 | 25.0 |
| | Total | 28 | 100 |
| Gender | Male | 5 | 17.9 |
| | Female | 23 | 82.1 |
| | Total | 28 | 100 |
| Years of Teaching English | 1-5 years | 18 | 64.3 |
| | 6-10 years | 6 | 21.4 |
| | Over 10 years | 4 | 14.3 |
| | Total | 28 | 100 |

**Rater trainer:** A professional rater who was experienced enough to run a rater training program was the workshop leader. The rater was an educated American native-speaker who was an instructor and an authorized scorer for the TOEFL iBT. With the help of a university professor of applied linguistics, she designed the materials for the training program and selected the rating rubric for workshop activities. In view of her educational background and academic experience, she had been actively involved in "implementing stringent rubrics developed by ETS for assessment of tests of native-speaker writing skills and of EFL/ESL writing and speaking proficiency." She also "earned certification and demonstrated ongoing expertise through regular calibration" (personal communication, April 28, 2011). The specialist in applied linguistics developed a plan for the instructional sessions and then the rater trainer defined the details of agenda on the basis of the workshop plan.

## 4.2 Instruments

The teachers went through an intensive rater-training course with a training package consisting of speaking assessment tasks and training program plan.

**Speaking assessment tasks:** As many as 10 recorded speaking monologs of EFL learners were given to the teacher raters to rate before and after the training program. The teachers rated the 10 monologs on a 4-point Likert scale: "1=weak," "2=fair," "3=good," and "4=very good." To discover the EFL teachers' criteria for rating L2 speaking ability, a separate section was employed to ask teachers to name as many criteria as they felt applicable for rating speaking. After rating the tasks, they mentioned the criteria they applied for rating and specified the degree of importance of each criterion on a 3-point Likert scale, ranging from "1=of little importance" to "3=very important."

**Training program:** The program consisted of four main parts: significance of speaking in L2 learning, construct of speaking, specification of speaking tests and speaking ratings, and significance of using criteria for rating. The aim was to help participants recognize the significance of assessing speaking ability with the application of rating criteria required for more effective and consistent rating. Besides, the rating rubric, along with the corresponding level descriptors, was described in the program so that the teacher raters could understand assessment criteria and draw on them for speaking assessment. An adapted version of the rating scale of ETS (2001) and a scale developed by Phillips (2008) were used as the rubric in the training program. The ETS scale is based on the theory of communicative language ability (i.e. functional, sociolinguistic, discourse, and linguistic). The rating scale developed by Phillips to score speaking tasks includes seven criteria, namely "answer to question," "comprehensibility," "organization," "fluency," "pronunciation," "grammar," and "vocabulary." Thus, the workshop rating rubric encompassed resources ranging from linguistic resources (i.e. vocabulary, grammar, pronunciation) to pragmatic aspects of language use (i.e. fluency, organization, comprehension, and thematic development).

### 4.3 Data Collection Procedure

Prior to rater training, the 10 speaking tasks were rated by each of the 28 EFL teachers. The teachers also rated the same tasks after attending the training program. Teachers' criteria in rating L2

speaking ability were explored at two phases of pre- and post-training. To check the initial inter-rater reliability of raters on assigning scores to 10 monologs, an acceptable inter-rater reliability index of 0.84 was calculated via intraclass correlation. The teacher raters attended a two-day training program, lasting six hours with one-week time interval between the first and second sessions. Through the training sessions, they received instruction and got acquainted with speaking rating criteria. They engaged in rating practice to rate monologic task types. The rating rubric was established to guide the teachers to abstract the key features of performance at different band levels. Negotiations and interactions among the teachers who had diverse opinions became the "learning moments." Meanwhile, the teachers applied a set of criteria to assign scores to each speaking task through individual and collaborative ratings. They were introduced to the selected rating criteria for scoring and then practiced using the rating scale to score audiotaped monologs. After the training program, the teachers assessed the 10 monologs again and specified their criteria for speaking assessment.

### 4.4 Data Analysis

The analyses of the data included both descriptive (frequency and mean score) and inferential (chi-square and t-test) statistics for speaking rating criteria and their significance. The teachers' speaking rating criteria were then grouped into 10 categories. Categorization was made in view of the common characteristics of related criteria. For instance, a single category named "topic management" was used to integrate the three criteria of topic relevance, topic coverage, and topic mastery specified by the teachers. The teachers' criteria were subsequently categorized into pre- and post-training criteria and then their frequencies were calculated. Chi-square was used to calculate differences in the frequencies. Then differences in the significance the raters assigned to each of the criteria before and after training were investigated through t-tests.

# 5. Results

## 5.1 Pre- and Post-training Criteria for Rating Speaking

The first research question was: What criteria do non-native EFL teachers use to rate the speaking performance of EFL learners before and after the speaking rater training? To investigate the question, the criteria the EFL teachers mentioned for rating speaking were analyzed. This analysis was based on the coding system of the identification of key speaking components. Based on the empirically derived criteria, 10 categories of speaking rating criteria are (1) fluency, (2) grammatical accuracy, (3) vocabulary, (4) pronunciation, (5) comprehension, (6) topic management, (7) rate of speech, (8) affective variables, (9) organization, and (10) function. In this study, the speaking rating criteria were grouped into three categories: (1) common criteria the EFL teachers used both before and after training, (2) pre-training-specific criteria, and (3) post-training-specific criteria (Table 2). Regarding the common criteria, 10 categories were discovered for rating speaking both before and after training. The specific criteria mentioned by the teachers before training fell into three sub-categories: persuasiveness, emotion, and interaction. Finally, three sub-categories were derived from the EFL teachers' criteria after training: intelligibility, self-confidence, and topic development.

**Table 2:** Criteria used by the EFL teachers for speaking rating before and after training

| **Major categories** | **Sub-categories** |
|---|---|
| 1. Fluency | Naturalness and appropriateness |
| 2. Grammatical accuracy | Range in grammatical structures<br>Register |
| 3. Vocabulary | Lexical resources<br>Lexical maturity |
| 4. Pronunciation | Prosodic features<br>Voice quality |
| 5. Comprehension | Audience awareness<br>Intelligibility** (post-training-specific criteria) |

| | | |
|---|---|---|
| 6. | Topic management | Topic relevance<br>Topic coverage<br>Topic mastery |
| 7. | Rate of speech | Delivery |
| 8. | Affective variables | Creativity/Persuasiveness* (pre-training-specific criteria)<br>Emotion/Engagement/Rapport* (pre-training-specific criteria)<br>Self-confidence** (post-training-specific criteria) |
| 9. | Organization | Topic development** (post-training-specific criteria)<br>Coherence and Cohesion<br>Preparation |
| 10. | Function | Interaction and conversation management* (pre-training-specific criteria) |

*Note: *pre-training-specific criteria; **post-training-specific criteria.*

### *5.1.1 Description of Common Criteria*

The criteria mentioned by the EFL teachers both before and after training in terms of their sub-components are described below:

**(1) Fluency:** It included those aspects of spoken language that contribute to the smoothness and natural flow of ideas (i.e. initiation, maintenance, stalling) without too much pauses and hesitations.

**(2) Grammatical accuracy:** It was related to syntax and morphology with the corresponding features of range, variety, appropriateness, and register (e.g. level of formality and politeness).

**(3) Vocabulary:** It ranged from those aspects associated with linguistic maturity (i.e. repertoire of words, variety, and appropriateness) to the categories of lexical choices (i.e. authenticity, naturalness, and simplicity vs. complexity).

**(4) Pronunciation:** It encompassed the prosodic features of spoken production (e.g. stress, rhythm, and intonation) and voice quality (e.g. voice volume and tone).

**(5) Comprehension:** It was based on the degree to which the speech could be understood by the listener.

**(6) Topic management:** This category was made up of three aspects: topic relevance, topic coverage, and topic mastery. The

ability of respondents to directly address the question while being focused on the topic (i.e. relevance) and to the point (i.e. lack of redundancy) referred to the first component of topic management. Also, topic coverage consisted of the adequacy of the answer (i.e. not missing any point) and the adequacy of elaboration and reasoning (e.g., using examples/details, providing proofs, and making claims). The last component was the mastery of the topic (i.e. knowledge and general range/content).

**(7) Rate of speech:** A salient feature of a fluid speech is to have an acceptable speed that was specified by the teachers as a distinct aspect along with other attributes of speaking construct.

**(8) Affective variables:** It consisted of the criteria related to the consideration of the interlocutor (i.e. persuasiveness, emotion, and engagement) and confidence in speech delivery on the part of the speaker.

**(9) Organization:** It was composed of the more general criterion of topic development and the more specific criteria of transitions, coherence, and cohesion.

**(10) Function:** The overall features of communication were categorized in terms of interaction and participation.

As evident from the above list, the linguistic resources were mentioned by the teachers as salient criteria when assessing EFL learners' speaking ability both before and after training.

### *5.1.2 Description of Pre-training-specific Criteria*

The pre-training-specific criteria were mentioned by the teachers before training. They tended to fall mostly within affective variables and interactive communication skills, as described below:

**(1) Creativity/Persuasiveness:** By this criterion, the teachers meant the ability to keep the topic new and interesting while being persuasive enough. In other words, the consideration of the interlocutor's interest to keep his/her attention was regarded as one of the affective features for rating L2 speaking ability.

**(2) Emotion/Engagement/Rapport:** The affective variables in terms of interpersonal skills (e.g., engagement/rapport) and nonverbal behavior (e.g. eye contact, body language, and

gestures) with respect to the conveyance of a message were also considered by the teachers in their speaking ratings.

**(3) Interaction:** The adequacy of participation (e.g. the length of the person's contribution) to effectively get the meaning across with the least effort (e.g. communicative effectiveness) while observing the native speakers' norms (e.g. awareness of cultural differences) were all within the category of communicative functions.

The specific pre-training criteria show that the teachers emphasized the overall communicative nature of speaking in view of its affective and interactive features. However, there was a shift in the criteria used for rating speaking after training.

### *5.1.3 Description of Post-training-specific Criteria*

The criteria mentioned by the EFL teachers after training suggest that they focused more on higher-level components for rating speaking, such as:

**(1) Intelligibility:** The automaticity in speech production with respect to intelligibility (i.e. voice projection and articulation) while being able to keep going comprehensibly was considered as significant by the teachers after attending the speaking rating program.

**(2) Self-confidence:** The affective variable of self-confidence in delivering the speech was also given some weight by the teachers.

**(3) Topic development:** The coherence of expression (i.e. how to start, continue, and end) and preparation in terms of clarity of ideas were mentioned as one of the key characteristics of a well-developed speech.

As the criteria show, an increasing understanding of the higher-order aspects of the speaking construct, such as comprehension and organization, resulting from attending the training program helped the trainees become aware of the hidden aspects of speaking and to look for more higher-order components.

### *5.1.4 Frequency and Significance of Pre- and Post-training Criteria for Rating Speaking*

The second research question of the study was concerned with the difference in the frequency and significance of criteria non-native EFL teachers use to rate the speaking performance of EFL learners before and after speaking rater training. To address the question, the teachers were asked to name those criteria that they considered as the salient features of L2 speaking performance for assessment purposes. Table 3 presents the teachers' rating criteria by frequency prior and subsequent to training. Moreover, the significance teachers attributed to each of the criteria is given in terms of the mean scores on a 3-point Likert scale.

**Table 3:** Descriptive statistics of frequency and significance of pre- and post-training criteria for rating speaking

| Speaking Rating Criteria | Frequency (N=28) | | Percentage (%) | | Mean (3-point Likert scale) | |
|---|---|---|---|---|---|---|
| | pre-training criteria | post-training criteria | pre-training criteria | post-training criteria | pre-training criteria | post-training criteria |
| Fluency | 17 | 27 | 11.1% | 14.8% | 1.71 | 2.75 |
| Grammatical accuracy | 23 | 27 | 15.03% | 14.8% | 2.18 | 2.57 |
| Vocabulary | 20 | 23 | 13.07% | 12.6% | 1.82 | 1.89 |
| Pronunciation | 26 | 26 | 17% | 14.3% | 2.36 | 1.96 |
| Comprehension | 8 | 27 | 5.2% | 14.8% | .86 | 2.86 |
| Topic management | 18 | 19 | 11.8% | 10.4% | 1.68 | 1.89 |
| Rate of speech | 10 | 4 | 6.5% | 2.2% | .86 | .29 |
| Affective variables | 11 | 2 | 7.2% | 1.1% | .93 | .18 |
| Organization | 17 | 26 | 11.1% | 14.3% | 1.64 | 2.36 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Function | 3 | 1 | 2% | 0.55% | .29 | .11 |
| Total | 153 | 182 | 100% | 100% | 1.43 | 1.68 |

The majority of the observed pre-training criteria were linguistic resources, i.e. pronunciation (frequency: 26, percentage: 17%), grammatical accuracy (frequency: 23, percentage: 15.03%), and vocabulary use (frequency: 20, percentage: 13.07%). Regarding the salient aspects of speaking assessment, the teachers also mentioned those aspects of language use involving topic management (frequency: 18, percentage: 11.8%), fluency (frequency: 17, percentage: 11.1%), and organization (frequency: 17, percentage: 11.1%). The least-frequently mentioned criteria by the teachers before attending the program were comprehension (frequency: 8, percentage: 5.2%) and communicative effectiveness (frequency: 3, percentage: 2%). As the above list of criteria indicates, variability in the significance given to different criteria is evident from the weight given to each criterion. The most significant criteria were pronunciation (M=2.36) and grammatical accuracy (M=2.18) and the least significant ones included function (M=.29), comprehension (M=.86), and rate of speech (M=.86).

After training, there was a remarkable change in both frequency and significance of the derived criteria with reference to the higher-order components of speech such as fluency (frequency: 27, percentage: 14.8%), comprehension (frequency: 27, percentage: 14.8%), and organization (frequency: 26, percentage: 14.3%). This suggests that there was a change in post-training criteria with the highest mean scores for comprehension (M=2.86), fluency (M=2.75), and organization (M=2.36). By contrast, there was a sharp decline in the significance of two criteria: rate of speech (M=.29) and affective variables (M=.18). After training, language components such as grammar (frequency: 27, percentage: 14.8%), pronunciation (frequency: 26, percentage: 14.3%), and vocabulary (frequency: 23, percentage: 12.6%) were still among those linguistic resources with a high degree of frequency. However, it is important to note that there was a decrease in the significance the teachers attributed to pronunciation from the mean score of 2.36 to that of

1.96. As a result, the frequencies of the 10 categories of speaking criteria indicate that the teachers were influenced by a set of criteria in their ratings with varying levels of significance.

Chi-square statistic was employed to investigate whether the differences in the frequencies of the criteria were statistically significant. As Table 4 shows, the difference between the obtained and the expected frequencies was large enough ($\chi^2$ =29.41, df=1, $p$<.01) to substantiate the conclusion that different rating criteria in terms of frequency influenced teachers' ratings before and subsequent to training.

**Table 4:** Chi-square test for the total speaking rating criteria

| | value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 29.41[a] | 1 | .000** |
| Continuity Correction[b] | 28.06 | 1 | .000** |
| N of Valid Cases 28 | | | |
| Table 2*2 (Continuity Correction value is reported.) | | | |

*Note:* ** significant at $p$<.01.

The differences in terms of the total frequency counts, of course, suggest that the EFL teachers in the present study experience a change in speaking criteria after training. However, none of the individual criteria showed any statistically significant difference from pre-training to post-training (Table 5), except for the grammatical accuracy ($\chi^2$ =4.770, df=1, $p$<.05).

**Table 5:** Chi-Square test for each of the speaking rating criteria

| Speaking Rating Criteria | | value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|---|
| Fluency | Pearson Chi-Square | 1.603[a] | 1 | .206 |
| | Continuity Correction[b] | .050 | 1 | .823 |
| Grammatical accuracy | Pearson Chi-Square | 4.770[c] | 1 | .029* |
| | Continuity Correction[b] | .730 | 1 | .393 |
| Vocabulary | Pearson Chi-Square | .219[d] | 1 | .640 |
| | Continuity Correction[b] | .000 | 1 | 1.000 |

| | | | | |
|---|---|---|---|---|
| Pronunciation | Pearson Chi-Square | .166[e] | 1 | .684 |
| | Continuity Correction[b] | .000 | 1 | 1.000 |
| Comprehension | Pearson Chi-Square | .415[f] | 1 | .520 |
| | Continuity Correction[b] | .000 | 1 | 1.000 |
| Topic management | Pearson Chi-Square | 2.274[g] | 1 | .132 |
| | Continuity Correction[b] | 1.179 | 1 | .278 |
| Rate of speech | Pearson Chi-Square | 3.137[h] | 1 | .077 |
| | Continuity Correction[b] | 1.458 | 1 | .227 |
| Affective variables | Pearson Chi-Square | 1.394[i] | 1 | .238 |
| | Continuity Correction[b] | .184 | 1 | .668 |
| Organization | Pearson Chi-Square | .104[i] | 1 | .747 |
| | Continuity Correction[b] | .000 | 1 | 1.000 |
| Function | Pearson Chi-Square | .124[j] | 1 | .724 |
| | Continuity Correction[b] | .000 | 1 | 1.000 |
| N of Valid Cases 28 | | | | |

*Note:* * significant at $p<.05$.

Overall, the findings mirror the influence of a variety of factors. In reality, teachers had already been aware of linguistic and pragmatic aspects of speaking ability before attending the program; however, the hidden aspects of speaking construct were explored with respect to the significance of criteria the teachers assigned after training. To this end, a paired-sample *t*-test was run to compare the mean scores related to the significance of pre- and post-training criteria. The *t*-value is 2.73 with a probability of .01 (Table 6). This amount of *t*-value at 27 degrees of freedom is greater than the critical *t*-value, i.e. 2.05. Based on these results, it can be concluded that there was a significant difference between the mean scores of the importance given to pre- and post-training criteria. Therefore, training was found to have a significant effect on the EFL teachers' perception of significance of total criteria. However, a distinction must be made between statistical significance and meaningfulness of the findings. The effect size (Cohen, 1988) of the *t*-observed-value of 2.73 is .057. Based on the criteria developed by Cohen, an effect size value of .06 is of a moderate value.

**Table 6:** Paired samples t-test for the significance of the total speaking rating criteria

| | M | SD | SEM | 95% CI | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|
| TotalSig.PrCr-TotalSig.PostCr | .25 | .49 | .09 | [.06, .44] | 2.73 | 27 | .011* |

*Note:* * significant at $p<.05$.

A comparison of the difference in the significance of each criterion appears in Table 7. As can be seen, the paired samples *t*-tests indicate that there were significant differences between means of five rating criteria: fluency ($t=3.98$; df=27, $p<.01$), comprehension ($t=7.48$; df=27, $p<.01$), organization ($t=2.73$; df=27, $p<.05$), rate of speech ($t=-2.52$; df=27, $p<.05$), and affective variables ($t=-2.59$; df=27, $p<.05$). Results of the *t*-test with respect to other criteria show that the differences between the mean scores were not statistically significant. This finding show that, out of the 10 rating criteria, 5 criteria received differential significance as a result brought about by the rater training program.

**Table 7:** Paired samples t-tests for the significance of each of the speaking rating criteria

| Significance of Speaking Rating Criteria | M | SD | SEM | 95% CI | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|
| Fluency | 1.03 | 1.37 | .26 | [.50, 1.56] | 3.98 | 27 | .000** |
| Grammatical accuracy | .39 | 1.13 | .21 | [-.047, .83] | 1.83 | 27 | .078 |
| Vocabulary | .071 | 1.65 | .31 | [-.57, .71] | .229 | 27 | .82 |
| Pronunciation | -.39 | 1.22 | .23 | [-.86, .083] | -1.69 | 27 | .102 |
| Comprehension | 2.00 | 1.41 | -.26 | [1.45, 2.54] | 7.48 | 27 | .000** |
| Topic management | .21 | 1.61 | .30 | [-.41, .84] | .70 | 27 | .490 |
| Rate of speech | -.57 | 1.20 | .22 | [-1.03, -1.06] | -2.52 | 27 | .018* |
| Affective variables | -.75 | 1.53 | .28 | [-1.34, -.15] | -2.59 | 27 | .015* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Organization | .71 | 1.38 | .26 | [.17, 1.25] | 2.73 | 27 | .011* |
| Function | -.17 | 1.05 | .20 | [-.58, .23] | -.892 | 27 | .379 |

*Note:* * significant at $p<.05$; ** significant at $p<.01$.

## 6. Discussion

The first finding of the current research revealed variations in raters' rating criteria due to training and the second one showed differences in the importance attributed to individual rating criteria. The results indicate that a number of linguistic resources encompass the criteria for rating speaking. The criteria reported by the non-native EFL teachers both before and after training suggest that the commonly used categories for rating EFL learners' speaking ability include the structural features of general range (content), vocabulary range, grammatical accuracy, and pronunciation. In addition, the frequency counts show that a set of rating criteria, including "affective variables," "function," and "rate of speech" are among those that are not frequently used by the EFL teachers when assessing learners' speaking performance. Based on the findings of the present study, it appears that the EFL teachers' rating criteria for L2 speaking are compatible with those reported in the previous studies (e.g. Barnwell, 1989; Brown et al., 2005; Chalhoub-Deville, 1995; Galloway, 1980; Hadden, 1991; Iwashita et al., 2008; Plough et al., 2010; Zhang & Elder, 2011). As indicated by the frequency of speaking criteria, the teachers mentioned linguistic features more than other rating categories. This is in correspondence with the findings of Kim (2009), who found that teachers were more critically oriented toward certain linguistic features of spoken production in their ratings such as pronunciation, specific grammar use, and accuracy.

Following the training program, the teachers tended to highlight both pragmatic features of speaking (e.g. intelligibility and topic development) and structural categories of speech (e.g. accuracy, vocabulary, and pronunciation). The salient finding was that the contribution of other rating criteria such as fluency, organization, and comprehension increased after training. The teacher trainees in

this study listened to monologs and practiced scoring them while gaining the insight not to strictly adhere to the limited aspects of communication (e.g. grammar and pronunciation) in their final ratings. In consequence, the EFL teachers reconsidered their criteria to include those more in line with communicative speaking assessment (i.e. message focused criteria). Changes in the non-native EFL teachers' rating criteria involved a more comprehensive view of speaking ability including both linguistic and non-linguistic factors. In effect, the training program led to the rising importance of the higher-order criteria required for more consistent rating. More importantly, these higher-order components are communicative performance features that the teachers focused on as a result of training. This corresponds with the findings of Ang-Aw and Goh's (2011) study, showing that the trained raters placed greater focus on content and fluency in comparison to other rating criteria like as language accuracy. It is also compatible with the findings by Iwashita et al. (2008) and Plough et al. (2010). Their findings indicate that the application of rating criteria may be subject to variation by raters' different perceptions of good speaking performance. Likewise, he training program in the current study inspired such variation in the teachers' perceptions of good speaking performance.

In addition to variation in teachers' rating criteria for L2 speaking, the findings of the study indicate that the teachers attach various amounts of weight to the individual rating criteria. The teachers varied significantly in their perceived general importance for speaking criteria to rate EFL learners' speaking ability. Prior to training, they focused more on underlying language ability in terms of the significance they attributed to the structural features of speaking performance such as "pronunciation" and "grammar." By contrast, their focus was redirected toward "fluency," "comprehension," and "organization," for example, indicating that they gained a more profound understanding of the speaking construct. For instance, they did not lend much weight to "organization" before attending the program, whereas they seemed to be strongly affected by this criterion in their final ratings subsequent to the training. Furthermore, "comprehension" came to be considered one of the important speaking rating criteria due to

training because the problem with comprehensibility may arise from the linguistic resources and prosodic features (Pennington & Ellis, 2000). It stands to reason that temporal variables, such as speech rate, pause, and hesitation, are regarded not only as the main concern in assessment but also of direct relevance to effective communication (Griffiths, 1990, 1991). Therefore, the primary focus of raters who are comprehenders (Douglas, 1997) is on the intelligibility criteria in processing and evaluating L2 oral ability (Hahn, 2004). Thus, as Douglas (1997) points out, raters may evaluate speakers more favorably, as the teachers in this study did after training, when they normally listen to the speakers' performance and make sense of the message.

These differences in the importance given to any particular component of the speaking construct, as Zhang and Elder (2011) argue, may be related to a dichotomy between a weak and strong approach in performance-based settings. Teachers may display a narrow scoring focus when the test task is at the service of discrete-point measures, whereas in the strong approach the performance features that teachers focus on are directed on the evaluation of "hows" of students' performance. As evidenced by the number of criteria, the linguistic features (i.e. pronunciation, grammar, and vocabulary) were still among the most frequently used criteria for rating EFL learners' speaking ability. However, the teachers' criteria for rating speaking after the training program became more inclusive in terms of their significance across higher-order components. With training, this study reveals, the teachers became aware of the hidden aspects of the construct salient to consistent speaking rating. As previously suggested by Chalhoub-Deville (1995), these findings indicate a return to integrative, communicative tests and a retreat from discrete-point components for testing L2 speaking ability of EFL learners.

## 7. Conclusion

Language teaching involves assessment of learners. The decisions teachers need to make concerning learners' speaking performance involve choices about the best ways of assessment. Speaking assessment, with specific reference to teaching and learning in the

language classroom, requires a communicative view of language in which both structural and pragmatic aspects of language should be attended to. Equally, as part of the professional development, teachers should take part in teacher training programs to gain insight into the construct of speaking and assessment rubrics. The change in the speaking assessment beliefs of teachers that takes place in the training program leads them to grasp an understanding of the speaking construct.

Rating criteria channel the ways teachers perceive the construct in question and hence arrive at their ratings (Eckes, 2008; McNamara, 1996). The scores that strongly adhere to the limited range of speaking criteria are not desirable to be representative of learners' speaking performance. If teachers only focus their attention on discrete-point components, losing sight of the way the learners get a message across, there is a possibility of introducing negative rating washback. Overall, the training program in this study was effective in encouraging teachers to attend to macro-level, higher-order components when making a global judgment of speaking performance. Changes in the importance non-native EFL teachers assign to the criteria for rating L2 learners' speaking ability confirm the suitability of teacher training courses of this type. Thus, the observed changes after the training program need to be embedded in further teacher education programs.

Teachers' rating behavior, particularly with regard to a change in their applied rating criteria and more importantly the weight they assigned to each of the criteria, contributed significantly to the effect of the rater training program. Nevertheless, the following limitations in this study should be taken into consideration before making any generalization. First, to select teacher raters, a convenience, rather than random, sampling was used in the current research. Second, learners with the same L1 background and level of proficiency took oral tests. Finally, the non-native EFL teachers were asked to rate only a particular task type, monolog, before and subsequent to the training program with the application of holistic rating; however, as Lievens (2001) states, there is a need to retrain teachers who display a narrow focus in their ratings to use a bottom-up, analytic scoring.

Once again, as Douglas (1997) argues, more research is needed to know more about how EFL teachers arrive at their speaking ratings and how they assign the same scores for somewhat different reasons. It appears, then, that more studies of raters' use and perception of criteria for rating speaking before and after rating training need to be conducted. As a result, it remains to be investigated whether the speaking rating criteria derived in this study are transferable to other speaking test tasks and contexts.

## References

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

Bachman, L.F., Lynch, B.K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257.

Barnwell, D. (1989). 'Native' native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6*(2), 152-163.

Brindley, G. (1991). Defining language ability: The criteria for criteria. Unpublished manuscript.

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

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

Brown, H.D. (2004). *Language assessment: Principles and classroom practices*. White Plains, New York: Longman.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks*. ETS Research Report No. RR-05-05. Princeton, New Jersey: Educational Testing Service.

Caban, H.L. (2003). Rater bias in the speaking assessment of four L1 Japanese ESL. *Second Language Studies, 21*(2), 1-44.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Chuang, Y. (2009). Foreign language speaking assessment: Taiwanese college English teachers' scoring performance in the holistic and analytic rating methods. *The Asian EFL Journal Quarterly, 11*(1), 152-175.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic press.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations.* TOEFL Monograph Series Report No. 8. Princeton, New Jersey: Educational Testing Service.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221.

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

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*(3), 235-254.

Elder, C., Barkhuizen, G., Knoch, U., & Randow, J. V. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*(4), 347-368.

ETS (2001). *Information Bulletin for the Test of Spoken English*. TSE 2001-02. Princeton, New Jersey: Educational Testing Service. Online version of a bulletin available from http://www.toefl.org/tse/tseindx.html.

Fayer, J. M., & Krasinski, E. (1987). Native and non-native judgments of intelligibility and irritation. *Language Learning, 37*(3), 313-326.

Fulcher, G., Marquez Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321-344.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64*(4), 428-433.

Griffiths, R. (1990). Speech rate and NNS comprehension: A preliminary study in time-benefit analysis. *Language Learning, 40*(3), 311-336.

Griffiths, R. (1991). Pausological research in an L2 context: A rationale and review of selected studies. *Applied Linguistics*, *12*(4), 345-364.

Hadden, B. (1991). Teacher and non-teacher perceptions of second-language communication. *Language Learning*, *41*(1), 1-20.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201-223.

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.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*(3), 401-436.

Kenyon, D. M. (1992). *Development and use of rating scales in language testing*, remarks introducing symposium. Paper presented at the 14th annual Language Testing Research Colloquium, Vancouver, British Columbia, Canada.

Kim, H. J. (2005). *World Englishes and language testing: The influence of rater variability in the assessment process of English language oral proficiency* (Unpublished doctoral dissertation). Iowa University, Iowa.

Kim, H. G. (2010). Investigating the construct validity of a speaking performance test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *8*, 1-30.

Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*(2), 187-217.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*(2), 255-264.

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes*, *17*, 347-367.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415-437.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

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

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180.

MacIntyre, P. N. (1993). The importance and effectiveness of moderation training on reliability of teachers' assessment of ESL writing samples. Unpublished MA thesis, University of Melbourne.

McNamara, T. F. (1995). Modelling performance: Opening Pandora's box. *Applied Linguistics*, *16*(2), 159-179.

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

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, *28*(4), 483-508.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task based second language performance assessment. *Language Testing*, *19*(4), 395-418.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing*, *19*(2), 169-192.

O'Sullivan, B. (2002). Learner acquaintanceship and OPT pair-task performance. *Language Testing*, *19*(3), 277-295.

Pennington, M., & Ellis, N. (2000). Cantonese speakers' memory for English sentences with prosodic cues. *The Modern Language Journal*, *84*(3), 372-389.

Phillips, D. (2008). *Longman preparation course for the TOEFL test: iBT speaking*. New York: Pearson Education.

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.

Robinson, P. (2001). Task complexity, task difficulty and task production: Exploring interaction in a componential framework. *Applied Linguistics*, *22*(1), 27-57.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, *11*(2), 99-123.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, *76*(1), 27-33.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques.* Cambridge: Cambridge University Press.

Weigle, S. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305-323.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 85-106.

Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps?* (ETS Research Report No. RR-09-31). Princeton, New Jersey: Educational Testing Service.

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