http://dx.doi.org/10.20880/kler.2016.51.5.83

Comparing Native and Non-native Rater Assessments of Korean Oral Proficiency:

A FACETS analysis

Kim, Hyunah University of Toronto

I. Introduction

Assessments of proficiency are often necessary in order to measure student progress and to determine preparedness for further levels of study. Because they often carry high stakes implications, researchers have long stressed the care with which they should be constructed, emphasizing the importance of validity and reliability. Especially, tests that involve rater judgment may be particularly prone to biases. Among the four language skills, speaking is said to be most susceptible to raters' subjective judgment. For this reason, Lee et al. (2003) argued that what is the most important in speaking assessment is to secure raters, or teachers, who have been given sufficient professional training in oral proficiency and oral assessment so that the assessment can be carried out with high reliability (p. 228). Although the matter of reliability can be viewed as less important for less-informal speaking tests conducted in classroom environment, ensuring high inter-rater reliability is critical for educational decision making that impacts outcomes, such as level placement and advancement (Luoma, 2004: 176-179). Since raters' subjectivity cannot possibly be excluded from speaking assessment, rater variables should be substantially considered

and investigated. According to Yu (2011: 5-6), for the last few decades rater bias caused by different rater variables, such as linguistic background (L1), occupation, and rater training in speaking assessment has been an area of significant research interest in second language education. Among the many factors, the L1 of raters, that is, whether they are native speakers (NS) or non-native speakers (NNS) of the language, appears to be one of the most influential in judgments of oral language proficiency.

With the recent worldwide increase in the number of overall Korean language learners and NNS instructors of Korean language, the L1 of assessment raters has become an even more noteworthy area of interest. It might be said that many NNS teachers of Korean feel less confident when it comes to teaching speaking compared to other skills. They have reported general unease when assessing oral language, noting the difficulty of avoiding subjectivity (Baek & Yang, 2011; Lee et al., 2003). In other words, many NNS teachers of Korean view themselves as lacking in judgment as raters of oral proficiency and, therefore, not able to secure an adequate level of validity and reliability of the assessment.

Despite the immediateness and practicality of this issue, however, no relevant study on how raters' L1 variance affects ratings in speaking assessment has been conducted in the Korean language teaching context. Therefore, in order to contribute to testing the validity and reliability of speaking assessment by NNS raters of Korean oral proficiency, this study aims to empirically investigate any possible systematic differences in evaluation of NS and NNS raters on Korean speaking tests. In addition, it explores the effects of some common devices used to enhance inter-rater reliability, such as detailed rubrics and rater training, in alleviating differences, if any, stemming from L1 background.

Ⅱ. Literature Review

1. Rater variables in Korean speaking assessment

Until this point, not much literature can be found that investigates the systemic effects of rater variables in speaking assessment in the field of Korean education as second or foreign language (KSL/KFL). In one of the few, Lee (2013) employed FACETS analysis to examine rater bias in estimations of pronunciation on a simulated Korean speaking assessment. Twenty raters from varying majors and with varying lengths of teaching experience were asked to judge the pronunciation of 20 speaking samples without detailed rubrics or rater training session. The results revealed that there was a significant difference in severity among rater groups while the raters in all groups successfully maintained an acceptable level of intra-rater reliability. Specifically, the two groups that demonstrated the most severe rating patterns were those with less than 5 years of teaching experience and those who concentrated in phonology.

Kang & Ahn (2014), likewise, used FACETS analysis to investigate bias in speaking assessments of foreign speakers of Korean. In this case, rater groups were divided into two clear groups, with 12 professional raters and 12 non-professional raters. Participants were asked to evaluate 27 speaking samples from a simulated computer-based speaking test using both holistic scoring and analytic scoring (i.e. vocabulary, pronunciation, grammar, and discourse). The analysis revealed that the professional rater group demonstrated a higher intra-rater reliability and an appropriate level of severity, although both the professional and non-professional rater groups managed to retain proper intervals between scores. In addition, since a considerable number of the non-professional raters showed misfit or overfit ratings, Kang & Ahn emphasized the importance of appropriate rater

training for reliable language assessment.

Including the above-mentioned two studies, a total number of three studies[1] were all that could be found regarding rater variables in Korean speaking assessment, showing a severe lack of literature in the field. Furthermore, since all these studies were focused on the rater experience variable and no NNS raters were included in the experiments, any comparison of judgments between NS and NNS raters was impossible in the first place.

2. Native and non-native rater variable in speaking assessment

Unlike in the field of Korean language education, a great deal of research exploring the effects of raters' language backgrounds on speaking assessment has been conducted in the field of English as a second or foreign language. As one of such initial studies, Fayer & Krasinski (1987) compared the differences between judgments by 40 native speakers (English speakers) and 88 non-native speakers (Spanish speakers) on 7 English speaking samples produced by ESL college students in Puerto Rico. It was revealed that the NNS group gave significantly lower scores for linguistic form and reported annoyance more frequently, while the scores for intelligibility given by the two rater groups were similar. Therefore, the researchers argued that NNS teachers should focus less on "features that are not particularly annoying to the NS listeners." (p. 325)

A recent study by Baek & Yang (2011) targeting Korean EFL learners similarly found NNS raters to be more severe. In this study, three NS teachers and 5 NNS (Korean) teachers of English from Korean middle schools were asked to rate 18 speaking samples from a simulated speaking

1 In the third relevant study by Lee (2014), the effects of teaching experience were explored by dividing the raters into three groups: i) the non-teacher group, ii) the teacher group with less than 5 years of teaching experience, and iii) the teacher group with more than 5 years of teaching experience.

test. Rating patterns were investigated and analyzed using both FACETS analysis and one-on-one interviews. The analysis displayed that, whereas rating patterns acceptably coincided within the NS rater group, ratings by NNS teachers were problematic in terms of their low inter-rater reliability.

Some recent studies, however, have interpreted similar results in a contrasting point of view. In a study by Kim (2006), 30 NS (English) teachers and 30 NNS (Korean) teachers participated in both holistic and analytic ratings of Korean EFL learners. A one-way analysis of variance (ANOVA) determined that there were no statistically meaningful differences in holistic evaluation between NS and NS teacher groups, although the two groups "differed in some analytic ratings of the Korean students' speech samples such as in rate of speech (i.e. fluency), organization, and task fulfillment" (p. 115). Nevertheless, both groups showed a reasonable level of inter-rater reliability for both holistic and analytic ratings, and, moreover, the holistic ratings by NNS raters were even more balanced in that they were highly correlated to all of the analytic rating categories compared to those by NS raters. Thus, Kim proposed that the prejudice against NNS raters being less qualified than NS raters should be eliminated.

Going one step further than this equally qualified view between NS and NNS raters, some researchers have maintained that NNS raters might sometimes be considered more appropriate, depending on the expected wash-back effect of the assessment. Zhang & Elder (2010), for instance, found this to be the case during their research in the English learning context. They asked 19 NS and 20 NNS (Chinese) raters to evaluate an official English speaking test in China, without any preceding rater training, and to submit written comments to rationalize their own rating on each speaking sample. While a FACETS analysis showed similar holistic ratings between the two groups, noticeable differences were found in the rationalization of the ratings and the constructs defined by each group: "a

native speaker rater would be more likely to pick up and comment on features of interaction, whereas a non-native speaker would be more likely to focus on linguistic resources such as accuracy" (p. 44). Therefore, Zhang & Elder imply that, for a speaking assessment of which the anticipated wash-back effect is to improve accuracy, rating by NNSs can be even more useful than that by NSs.

Apart from the above-mentioned literature, a number of studies have investigated the effects of raters' linguistic backgrounds on speaking assessment in the ESL or EFL context (Kim, 2009; Yu, 2011; Stassenko et al., 2014; Wei & Llosa, 2015) and, in most of these studies, NNS raters have been generally considered to focus more on linguistic forms rather than communicative competence and to show higher severity. However, since no studies have dealt with linguistic background as rater variable in the field of Korean language education and, moreover, a contrasting finding of more-severe NNS raters, especially in holistic rating and analytic rating of accuracy, was detected in the process of conducting related research by the present researcher, an empirical study on the NS and NNS judgments in Korean speaking assessment is highly needed. In addition, no previous studies have discussed the effectiveness of devices, such as detailed rubrics and rater training, in controlling raters' linguistic background variable and, thus, enhancing inter-rater reliability.

Accordingly, this study addresses the following questions:

1) Are there any systematic differences in native and non-native rater assessments of Korean oral proficiency?
2) Can methods aimed at strengthening inter-rater reliability, such as detailed rubrics and rater training, effectively narrow gaps between the native and non-native rater groups?

Ⅲ. Method

1. Participants

A total number of nine raters, comprised of four NSs and five NNSs, participated in this study, all of whom were graduate students in the Korean Language Education program at Seoul National University. In addition, all participants had had some degree of teaching experience at the time of experience. The L1 of each NNS participant varied, with two Chinese, two Japanese, and one Sinhala. Their mean length of stay in Korea was 5.2 years. Ages of both groups of participants raged between 29 and 41 (with the average of 32), and their teaching experiences also ranged from one year to six years. Detailed information of each rater is provided in Table 1.

Table 1. Detailed information on each rater

Rater group	Rater	L1	Length of stay in Korea (year)	Teaching experience	Age	Status
Native Korean speakers (Kor)	K1	Korean	N/A	6	33	PhD coursework
	K2			2	41	PhD coursework
	K3			4	32	PhD candidate
	K4			1	29	MA coursework
Non-native Korean speakers (Non-Kor)	N1	Chinese	5	2	30	PhD coursework
	N2	Chinese	3	1	29	PhD coursework
	N3	Japanese	5	2	29	PhD candidate

N4	Japanese	8	1	32	PhD coursework
N5	Sinhala	5	4	32	PhD coursework

2. Materials

The materials to be evaluated in the present study were mock Korean speaking test samples obtained from five learners of Korean at the beginner level. The examinees are from various countries including the United States, England, Ireland, and China, with different lengths of stay in Korea ranging from 1 to 5 years. Their occupations also varied and included language teachers, a student, and an artist. Some of the examinees had learned Korean through a formal language program at an official language institute, while others had no Korean learning experience in an orthodox classroom setting.

The speaking test was administered in the form of a 'simulated oral proficiency interview', or 'semi-direct interview'. It is acknowledged that this type of assessment has been partly criticized in that it is more or less one-sided with no reaction or feedback given by the examining listener and, thus, does not reflect authentic communication. However, it is still widely used in a number of acknowledged oral proficiency tests such as TOEFL iBT, TEPS-Speaking, and ACTFL OPI, not only for economic reasons, but also because the interlocutor factor can be effectively controlled (Lee, 2012: 343-344). In order to control this interlocutor factor and to facilitate delayed rating instead of on-the-spot rating, the type of this semi-direct interview was adopted in the present study.

The speaking stimulus provided for the examinees was a simple Korean sentence: "What did you do during your last summer holidays in where? Tell us about your experience." Upon receiving a piece of paper with the

text above, each examinee was given three minutes to prepare his or her answer and was asked to speak for as long as 90 seconds, without any interaction with the administrator.

3. Instruments

The collected speaking samples were assessed through two separate rounds: 1) first-round rating and 2) second-round rating. For each round of rating, both holistic scoring and analytic scoring were employed. The rating instruments and measures for each round are separately described as below.

1) First-round rating

For the first-round rating, raters received no training. The holistic scoring for this round was administered on an eleven-point scale from 0 to 10. In order to examine naturally occurring systemic differences, if any, between NS and NNS rater groups, no rating categories or rating criteria were provided in advance.

In regard to the first-round analytic scoring, the rating categories (or the rating constructs) were identified based on an extensive literature, with particular attention given to the constructs that are currently used in major oral proficiency tests, and the recommendations from Park et al. (2012). As a result, it was found that most leading tests had utilized five to six rating categories and that the categories of each test were practically the same as those of other tests, although the names of each category were slightly different. Since raters may be overwhelmed by rubrics containing a large number of rating categories and, therefore, may be less apt to judge them in detail (Lee, 2012: 358), the number of rating categories for the analytic scoring in this study was limited to five.

The final rating categories identified for this research were accuracy, range, fluency, contents, and organization, as described in Table 2. Pronunciation, which is separately assessed in some speaking tests (e.g. ACTFL OPI), was included as a component of Accuracy. In addition, Interaction as a category was excluded due to limitations of the semi-interview test format.

Table 2. The rating categories for analytic scoring

Rating category	Details
Accuracy	Is the grammar use and the pronunciation accurate?
Range	Is the speech clear with a wide range of vocabulary and linguistic expressions?
Fluency	Is the speech naturally fluent and with confidence?
Contents	Does the speaker understand the question or stimulus and respond with adequate contents?
Organization	Is the speech organizational and well-structured?

Heeding the advice of Lee (2012) that raters are inconsistent when provided too many points of scale (p. 359), and considering there are five different categories to take into account, the analytic scoring for each rating category was based on a 5-point scale, ranging from 0 to 4.

2) Second-round rating

The second-round rating incorporated a pre-training session. Without rating criteria to refer to, raters cannot be expected to effectively exclude their subjectivity not only in holistic, but also in analytic scoring. In this regard, it is generally believed among researchers and practitioners that providing verbal explanations for each point of scale that can be used as a basis of judgment (i.e. rating rubrics) is a useful

device to minimize rater bias in language performance assessment (Lee, 2012: 354). Therefore, the raters in both groups, after the first-round rating, were given detailed rubrics for both holistic and analytic ratings, while maintaining the overall assessment framework.

First of all, the rubric for holistic scoring was designed by the researcher based on the proficiency guidelines released by American Council on Teaching Foreign Languages (ACTFL, 2012). Considering the fact that all the examinees in this study are at the beginner level, only the proficiency guidelines for Novice Low, Novice Mid, Novice High, and Intermediate Low were used.[2] The fact that the speech samples were obtained from one-sided, semi-direct interviews with a single stimulus (i.e. speaking task) was also taken into account in developing the rubric. Moreover, instead of providing explanations for all 11 points on the holistic rating scale, the rubric provided four broader descriptions ((a) point 0 to 1, (b) point 2 to 4, (c) point 5 to 7, and (d) point 8 to 10) allowing raters some degree of leeway within these groups.

The rubric for each rating category of analytic scoring was designed by the researcher following the guidelines proposed by the Common European Framework of Reference for Languages (CEFR), the rubrics used in a well-known institute of Korean language, and those developed by Park et al. (2012).

4. Procedures

The oral rating portion of the experiment was carried out for about 120 minutes with all 9 participants present in a university classroom in Seoul.

2 The original guideline by ACTFL is divided into 10 levels: Superior, Advanced High, Advanced Mid, Advanced Low, Intermediate High, Intermediate Mid, Intermediate Low, Novice High, Novice Mid, Novice Low.

The overall procedure of the experiment is shown in Figure 1.

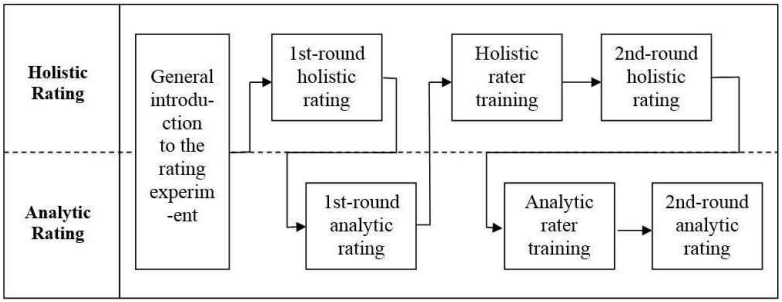


Figure 1. Procedure of rating experiment

The rating training sessions proceeded in the following order: (1) detailed rubrics were provided and the raters were given some time to become well-acquainted with the descriptions; (2) a Q&A session for clarification of the descriptions and the scales was held; (3) a mock rating session of Korean speech samples was conducted; (4) rater adjustment through discussions of the rating results between the raters and the researcher took place.

5. Analysis

In order to examine the characteristics and biases of each rater and rater group, FACETS program version 3.71.4 was utilized. The multi-facet Rasch measurement model (on which the FACETS program is based), an analyzing tool that has been most widely used to verify the validity and reliability of language performance assessment, stochastically traces the rating tendency of rater severity, rating category, or evaluation task (Kang & Ahn, 2014; Baek & Yang, 2011; Lee, 2014a; Lee, 2014b; Lee, 2013; Jang & Shin, 2009; Zhang & Elder, 2010).

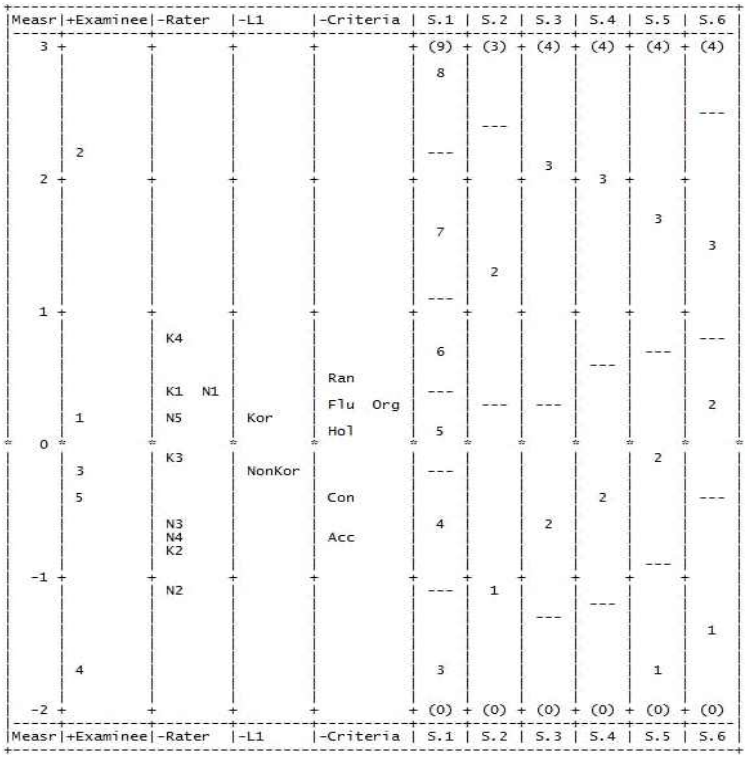
Particularly, Jang & Shin (2009) argued that FACETS is a valuable tool in providing bias indices of interaction between specific rater groups (e.g. NS and NNS rater groups) and examinees, evaluation tasks, or rating categories. In this study, four facets including examinees, rater severity, rater's L1, and rating categories were utilized for the analysis.

IV. Results

In order to examine the rating patterns of each rater and rater group in each round of rating sessions, a FACETS analysis was employed. The results are discussed in terms of 1) all facets vertical rulers, 2) rater severity, and 3) rater consistency.

1. All facets vertical rulers

The FACETS program produces a map called 'all facets vertical rulers' to exhibit the overall information on each facet on a logit scale. Figure 2 presents the comprehensive map of logit values of four facets (examinees, rater severity, rater's L1, and rating categories) for the first-round rating before the rater training. In the first column, the equal interval scale with a 'logit' unit is displayed to enable to compare the estimated measures within and between the facets.



S.1 Hol=Holistic, S.2 Acc=Accuracy, S.3 Ran=Range, S.4 Flu=Fluency, S.5 Con=Contents, S.6 Org=Organization

Figure 2. All facets vertical rulers (1st-round rating)

The second column, titled '+Examinee', depicts the oral proficiency of each examinee. The higher an examinee is located, the higher the examinee's proficiency level. The analysis shows that the examinee 2 received the highest score with a logit value of 2.16 while the examinee 4 scored the lowest with a logit value of −1.73.

The third column, titled '−Rater', provides information on the severity of the raters in evaluating the speaking samples. The higher a rater is located on the map, the more severe that rater is. Therefore, it can be interpreted

that one of NSs, K4 was the most severe (0.85 logit), while one of the NNSs, N2 was the most lenient (−1.14 logit). Considering the argument of Shin (2001) that this 2−logit spread in rater severity is converted into 40% difference in proficiency level received by examinees (p. 256), the discrepancy between K4 and N2 is not negligible.

The fourth column, titled '−L1', visually presents information on the severity of the rater groups in assessment of Korean oral proficiency. Similar to the third column, the higher a rater group is located on the map, the more rigorous the rater group is. The severity measure NS rater group (Kor) was 0.22 logit, while that of NNS rater group (NonKor) was −0.22 logit. The analysis suggests that the NS group was slightly more severe compared to the NNS group.

The fifth column, titled '−Criteria', illustrates rating scale difficulty of each rating category. The higher a rating category is located, the more severe the raters were on that category and the more difficult it was for the examinees to receive high scores. In this analysis, the severity on Holistic rating was moderate (0.12 logit), while the raters showed different levels of severity on each category of analytic rating in the order of Range (0.46 logit), Organization (0.28 logit), Fluency (0.26 logit), Contents (−0.40 logit), and Accuracy (−0.71 logit). In other words, Range was the most challenging for the examinees to gain high scores and Accuracy was the least challenging.

A series of columns towards the right end of the map ('S.1' column through 'S.6' column) provides information about the score distributions given to the examinees for each rating category. Although each point in each category was distributed in balance in general, the rating results tend to be pushed into the middle part of scale rather than the extremes.

Figure 3 presents parts of the 'all facts vertical rulers'maps for both 1st−round and 2nd−round rating in order to feature the change after

providing the scoring rubrics and the rater training. The comparison illustrates that, while maintaining the order of the examinees' proficiency level, the gaps of measurements among examinees became wider in the 2nd-round rating, better differentiating the ability of each examinee. As for the rater severity, most raters were placed on the lenient end, as seen by their location in the lower part of the third column (-Rater) in the map compared to the 1st-round rating. In addition, the gap of severity between the two rater groups narrowed from 0.44 logit to 0.28 logit, indicating a slightly higher inter-group reliability in the 2nd-round rating.

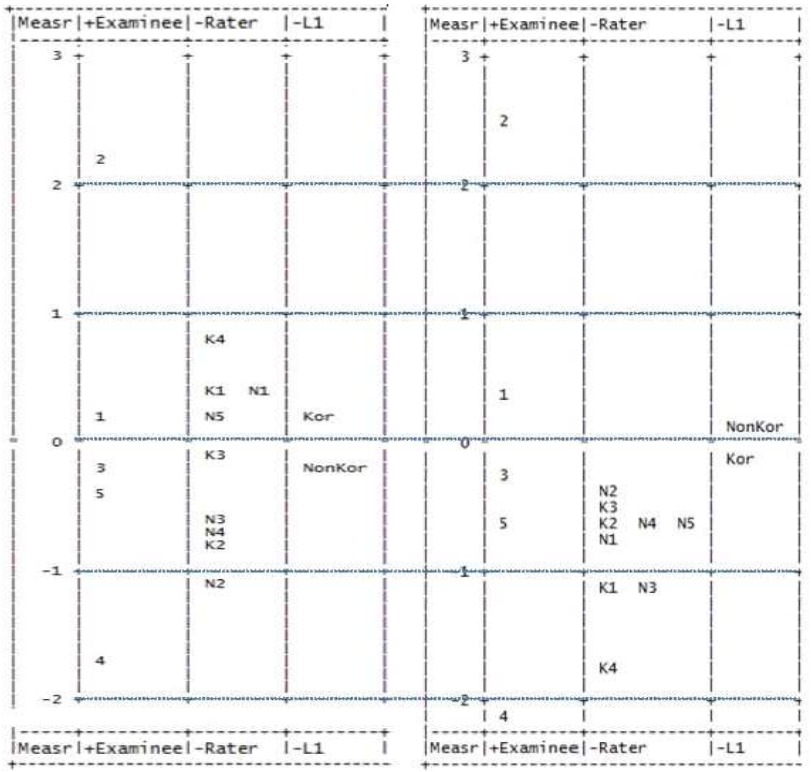


Figure 3. All facets vertical rulers (1st-round versus 2nd-round rating)

2. Rater severity

In order to examine whether the NS and NNS rater groups differ in severity in unguided rating of Korean oral proficiency, the measurement report by rater group for the 1st-round rating should be referred to. Table 3 shows the FACETS output analysis for group, anchoring the 'rater's L1' facet. Table 3 reveals that the fair average value of scores given by the NS group is 2.41, and that given by the NNS group is 2.73. Although, at first glance, this difference seems trivial, the high separation value of 2.53 and the high reliability of .86 indicate that the NS and NNS rater groups have different severity in their rating.[3] Furthermore, the significant chi-square value of 7.4 (df = 1, p = .01) also suggests significant difference in severity between the two rater groups.

Table 3. Measurement report by rater group (1st-round rating)

Rater group	Fair (M) Average	Severity measure	Model S.E.	Infit		Outfit	
				MnSq	Zstd	MnSq	Zstd
NS	2.41	.22	.12	1.01	.0	.98	-.1
NNS	2.73	-.22	.11	1.04	.3	1.01	.1
Separation 2.53, Strata 3.71, Reliability .86 Fixed (all same) chi-square: 7.4, d.f.: 1, Significance (probability): .01							

Moreover, whether the two rater groups differ in severity in any specific rating category (holistic rating or any category in analytic rating) was investigated. Figure 4 shows the severity measures (in logits) in holistic rating and in five rating categories for analytic rating by rater group, indicating that the two rater groups show disparate

3 The 'reliability' index is a different notion than the 'inter-rater reliability' index that is commonly used in language assessment studies. According to Jang & Shin (2009: 83), the lower the reliability index is, the more similar level of severity between raters or rater groups.

rating patterns in terms of severity in some rating categories. In general, the raters in the NS group (Kor) were more severe, with the logit values between −0.04 and 0.51, while the NNS group (NonKor) was comparatively lenient with the logit values between −0.43 and 0.00. This severity of the NS group was fairly consistent in both holistic and analytic rating, with the exception of the Fluency category of analytic rating. Specifically speaking, the more severe characteristic of NS raters was especially salient in holistic rating and in Accuracy and Range category of analytic rating.

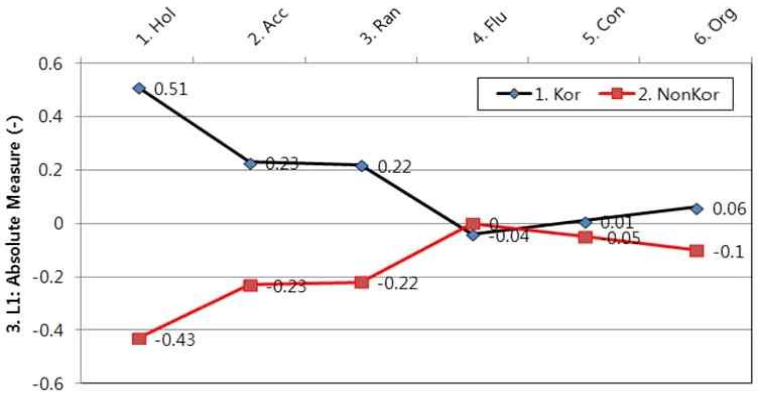


Figure 4. Rater severity of each category by rater group (1st-round rating)

Another analysis was carried out in order to identify any change in rating patterns between the NS and NNS rater groups after applying devices to enhance inter-rater reliability, that is, detailed rubrics and rater training. Table 4 shows the output from the group measurement report yielded by the FACETS analysis by rater group in the 2nd-round rating. In the 2nd-round rating, as can be seen, the fair averages given by the NS and NNS rater groups were 3.42 and 3.27, respectively,

which were somewhat higher than 2.41 and 2.73 from the 1st-round rating. What is particularly noteworthy is that the NNS rater group, who was more severe in the 1st-round rating, showed a slightly higher level of severity compared to the NS rater group in the 2nd-round rating. However, the difference in severity between the two groups has decreased noticeably, with the separation value of 1.27 and the reliability of .62. The non-significant chi-square value ($p = .11$) also shows that the two rater groups did not considerably differ in their severity after providing detailed rubrics and rater training.

Table 4. Measurement report by rater group (2nd-round rating)

Rater group	Fair (M) Average	Severity measure	Model S.E.	Infit		Outfit	
				MnSq	Zstd	MnSq	Zstd
NS	3.42	-.14	.13	.97	-.1	1.00	.0
NNS	3.27	.14	.11	.93	-.5	.93	-.5
Separation 1.27, Strata 2.03, Reliability .62 Fixed (all same) chi-square: 2.6, d.f.: 1, Significance (probability): .11							

Furthermore, it is necessary to examine how the difference in severity for each rating category changed in the 2nd-round rating. Figure 5 displays the severity measures (in logits) in holistic rating as well as in five rating categories for analytic rating by rater group. With the detailed rubrics and rater training provided, the differences in severity for each rating category between the two rater groups were comparatively stabilized overall. Specifically speaking, in the holistic rating and in the Range category of the analytic rating, the wide discrepancy that appeared earlier in the 1st-round rating (0.94 logit and 0.44 logit, respectively) declined markedly to 0.27 and 0.13 logit in the 2nd-round rating. Yet, the difference in severity between the rater groups for the Accuracy category of analytic rating widened from 0.46

logit in the 1st-round rating to 1.08 logit in the 2nd-round rating, suggesting that there was a problem in the process of providing detailed rubrics and rater training with regard to the Accuracy category. It is also possible that the participating raters might have over-adjusted during the rater training session, as the NNS group, who was more lenient in the 1st-round rating, assessed the speaking samples more harshly in the 2nd-round rating except the Range and Fluency category in analytic rating, and vice versa.

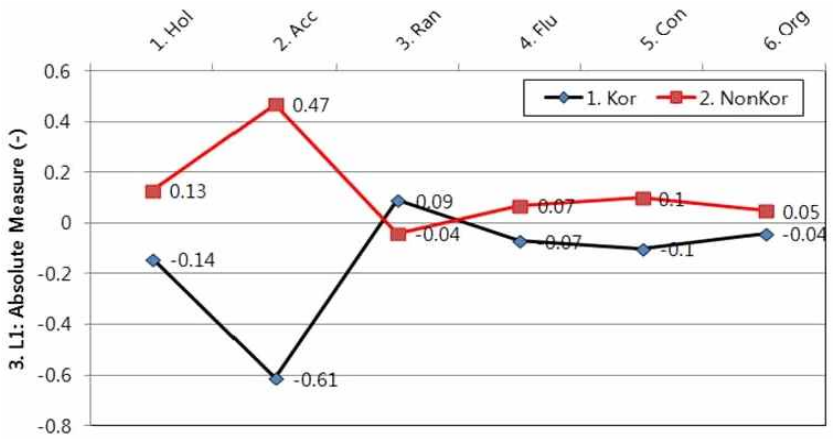


Figure 5. Rater severity of each category by rater group (2nd-round rating)

3. Rater consistency

Rater consistency, or intra-rater reliability, can be examined by 4 different indices produced by the FACETS analysis: infit mean square, infit z-standardization value, outift mean square, and outfit z-standardization value. According to Shin (2006; cited in Lee, 2013: 229), in a FACETS analysis of rating tasks or rating categories with a

small number of examinees and when their proficiency levels are not normally distributed, infit mean square is considered most appropriate and preferred index among the four indices. In the present study, as Lunz & Stahl (1990: 433) suggested, a rater's consistency is viewed as 'good fit' if the infit mean square value is between 0.5 and 1.5.[4] Accordingly, infit mean square values higher than 1.5 indicate a case of 'misfit', while values lower than 0.5 were signs of 'overfit'.[5]

In the 1st-round rating, as already shown in Table 3, the infit mean square value of the NS group was 1.01 and that of the NNS group was 1.04, indicating that both rater groups had already secured a very high level of rater consistency even before providing detailed rubrics and rater training, and this finding corresponds with Kim (2009). The rater consistency in the 2nd-round rating can be investigated in Table 4, according to which the high rater consistency was maintained in both groups with the infit mean square value of the NS and NNS group was 0.97 and 0.93, respectively.

Examining individual rater consistency, not as a rater group, however, reveals a few raters with unacceptable consistency level. Figure 6 displays the infit mean square values of each rater for the 1st- and 2nd-round rating. In the 1st-round rating, the infit mean square values of most raters fell into the acceptable range between 0.5 and 1.5, whereas a NS rater K4 and a NNS rater N2 were determined as 'misfit' with the value of 1.95 and 1.52, respectively. However, in the

4 As for the 'good fit' criteria, McNamara (1996) argued that "values in the range of approximately 0.75 to 1.3 are acceptable" (p. 173). However, a less strict measure proposed by Lunz & Stahl (1990) was applied in this study, since the participants were not professional raters at the time of experiment.

5 According to Kang & Ahn (2014), 'misfit' raters are inconsistent in the ratings, resulting in giving unexpectedly low scores for excellent speaking samples or unexpectedly high scores for inferior ones. 'Overfit' raters, on the other hand, lack in variability in their scoring.

2nd-round rating, all raters including these two raters performed properly in terms of rater consistency, with no single rater considered either misfit or overfit.

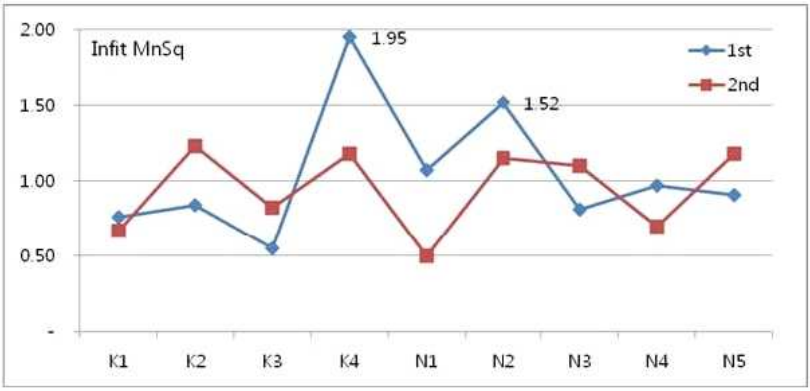


Figure 6. Rater consistency of each rater (1st- and 2nd-round rating)

V. Discussion and Conclusion

Motivated by the fact that no relevant research has been conducted in the Korean language teaching context, this study employed FACETS analysis to explore the differences between a rater group of native Korean speakers and that of non-native speakers in judging performance on a simulated Korean oral test of students at the beginner level. The primary purpose was to investigate the possibility of narrowing gaps between the two rater groups by means of some commonly used devices to enhance inter-rater reliability.

In order to answer the first research question, 'are there any systematic differences in native and non-native rater assessments of

Korean oral proficiency?', the rater characteristics of both groups were analyzed in terms of rater severity and rater consistency. First of all, the results suggested the NNS rater group assessed the speaking samples more leniently compared to the NS rater group and this tendency appeared to be more salient in holistic rating. As for analytic rating, the NNS rater group was more severe in the Accuracy and Range categories. These two categories of analytic rating, as mentioned above, are linked to the accuracy of grammar and pronunciation and the variety of vocabulary and expressions used, respectively, with a focus on linguistic features of speaking samples.

This 'severity of native speakers', especially in linguistic forms, coincides with the findings from the researcher's previous non-official observation, which was the very motive for this present study. Whereas, it contrasts to the widely known concept of 'severe non-native raters', especially in these linguistic-form-related categories, which was reported by previous research conducted in the ESL/EFL context (e.g. Fayer & Krasinski, 1987; Kim, 2006; Shin, 2001). Since the previous relevant research related mainly to raters of English oral proficiency, the unique results of the present study may possibly be attributed to the cultural differences among a certain group of countries (e.g. East Asian countries such as China, Japan, and Korea). As a matter of fact, Brown (1995), the only relevant study that could be found outside the ESL/EFL context, reported that native Japanese speakers showed a higher level of severity than non-native speakers in the assessment of Japanese oral test for tour guides. As for rater consistency, no significant differences between the two groups were detected, suggesting that both NS and NNS raters are able to secure an acceptable level of rater consistency regardless of the rater's L1 factor.

The second research question was: Can methods aimed at

strengthening inter-rater reliability, such as detailed rubrics and rater training, effectively narrow gaps between the native and non-native rater groups? The differences in rater severity between the two rater groups in the 1st-round rating were reconciled after applying the devices to enhance inter-rater reliability (specifically, detailed rubrics and rater training), implying the effectiveness of those devices. This finding is certainly meaningful because it opens up the possibility that, although intrinsic differences in rating severity between NS and NNS teachers of Korean exist, these differences can be controlled to some extent through clear and detailed rubrics and rater training.

As for rater consistency, it was found that that the reliability-enhancing devices such as detailed rubrics and rater training are effective in securing high rater consistency. Although the necessity of additional devices was marginal with an already-high level of consistency in the 1st-round rating, even some misfit or overfit patterns of a few individual raters were successfully eliminated in the 2nd-round rating, suggesting the high value of detailed rubrics and rater training. However, the findings suggest that rater training sessions should be planned and administrated more elaborately in order to guide the raters to not over-adjust.

The present study empirically compared NS and NNS rater assessments of oral proficiency in an effort to respond to the dearth of such studies in the KFL context and has shown that some differences in severity exist between the two rater groups. That said, these discrepancies can be addressed through the implementation of detailed rubrics and rater training. Nevertheless, these conclusions must be seen as very tentative given the following limitations. First of all, the numbers of participating raters and examinees in this study were very limited and not randomly selected. Particularly, since the most severe

rater K4 and the most lenient N2 were the least experienced in teaching Korean, the 'experience' factor, which was an uncontrolled variable in this study, should be taken into consideration for further studies. Furthermore, the most distinctive conclusion in this study – 'lenient non-native speakers' and 'severe native speakers' – needs continued investigation in order to verify whether this tendency exists prevalently in certain language communities.

The role of non-native teachers or raters is becoming more and more important with the increasing demands for Korean language education, since most of these demands exist in countries other than in Korea. In terms of rater consistency, as many studies including the present have already found, non-native raters seem as qualified as native speaking raters. Once more research is carried out on the difference in rater severity between NS and NNS raters and once more effective devices to enhance the inter-rater reliability are developed, many more proficient non-native language teachers will be able to assess their students' oral performance with greater confidence.

Submitted: 2016.10.31.
First revision received: 2016.12.09.
Accepted: 2016.12.09.

REFERENCES

Baek, H., & Yang, B. (2011). A study on middle school English teachers' rating patterns of speaking test. *Journal of Language Sciences*, 18, 77-99.

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

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

Jang, S. Y., & Shin, D. Y. (2009). *FACETS program for language education assessment research: basic*. Seoul: Global Contents.

Kang, S., & Ahn, H. (2014). A study of Korean raters' characteristics for foreign speakers' Korean oral performance. *Bilingual Research*, 55, 1-29.

Kim, H. J. (2006). Rater reliability in L2 oral proficiency tests. *English Teaching*, 61, 105-118.

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

Lee, H. (2013). Study of raters' rating patterns on a pronunciation criteria in speaking assessment - focusing on rater experience and major. *Teaching Korean as a Foreign Language*, 39, 213-245.

Lee, H. (2014). An investigation of the differences between novice teachers, experienced teachers and non-teacher native speakers in evaluation of Korean language learners' speech. *Journal of Korean Language Education*, 25, 163-188.

Lee, W. (2012). *A guide to English language testing*. Seoul: Moonjin Media.

Lee, Y. S., Lee, W. K., & Shin, D. Y. (2003). *Understanding language assessment*. Seoul: Seoul National University Press.

Lee, Y. S. (2014a). An investigation into the native raters' scoring of English writing assessment using the new Facets of many-facet Rasch

measurement. *English Language & Literature Teaching*, 20, 475-496.

Lee, Y. S. (2014b). A validation of the scoring of KFL writing assessment based on the many-facet Rasch measurement model. *Foreign Language Education*, 21, 355-375.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions*, 13, 425-444.

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

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

Park, D. H., Kim, Y. M., Kim, H. J., Shin, D. I., Woo, C. H., Lee, Y. S., Cho, S. J., & Ji, H. S. (2012). *Question type development for CBT/IBT-based Korean proficiency test*. National Institute for International Education.

Shin, D. I. (2001). Exploring rating patters with Rasch measurement techniques: Implications for training. *Foreign Language Education*, 8, 249-272.

Shin, D. I. (2006). *English language testing in Korea II: Speaking test*. Hankook Munhwasa.

Stassenko, I., Skopinskaja, L., & Liiv, S. (2014). Investigating cultural variability in rater judgements of oral proficiency interviews. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 10, 269-281.

Wei, J., & Llosa, L. (2015). Investigating Differences Between American and Indian Raters in Assessing TOEFL iBT Speaking Tasks. *Language Assessment Quarterly*,12,283-304.

Yu, K. A. (2011). *Effects of nonnative English-speaking raters' oral proficiency levels on English speaking test ratings and rating processes*. Ph. D. dissertation. Seoul: Ewha Woman's University.

Zhang, Y., & Elder, C. (2010). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?. *Language Testing*, 28, 31-50.

ABSTRACT

Comparing Native and Non-native Rater Assessments of Korean Oral Proficiency:

A FACETS analysis

Kim, Hyunah

Korean as a foreign language has experienced significant growth rates in the past few decades due in part to the worldwide popularity of Korean pop and entertainment culture, as well as its strengthening economic status. This new demand has led to a significant increase in the number of non-native Korean-speaking teachers. Despite their competence, many of these instructors report gaps in their confidence and abilities when it comes to evaluating students. Rater biases stemming from features present or absent in their linguistic and cultural backgrounds are of particular concern and may be more prevalent in assessments or oral proficiency. Despite the fact that rater bias in oral assessment has been explored extensively in the English-as-a-Foreign-Language context, there exists no relevant study of rater bias on speaking assessments in the Korean-as-a-foreign-language context. This paper reports the findings of an empirical study, which investigated possible systematic differences in evaluation of native Korean-speaking and non-native raters on speaking samples of learners of Korean and the effects of devices commonly used to enhance inter-rater reliability, such as detailed rubrics and rater training. The data were derived from four native and five non-native teachers of Korean, who were asked to evaluate five speech samples using both holistic and analytic rating scales. In the first round, the participants evaluated the samples without any rubrics. After being provided with

detailed rubrics and rater training, the participants again engaged in the second round of rating using the same speech samples. The results, analyzed by many-facet Rasch measurement, revealed that the rating patterns of the two rater groups were significantly different in terms of severity, especially for holistic rating and accuracy and range features for analytic rating, while both groups maintained the acceptable level of rating consistency. The analysis also showed that rating criteria and rater training substantially settled the difference in rating severity between the two rater groups. The paper concludes with implications of the study on NNS raters' rating pattern and future directions for rater training for Korean oral proficiency test.

KEYWORDS Korean speaking assessment, rater variables, rating patterns, native and non-native speakers, NS and NNS, many-facet Rasch measurement