THE JOURNAL OF ASIA TEFL
Vol. 15, No. 4, Winter 2018, 1036-1050
http://dx.doi.org/10.18823/asiatefl.2018.15.4.10.1036

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Different Rating Behaviors between New and Experienced NESTs When Evaluating Korean English Learners' Speaking

Kyoung Rang Lee
Sejong University, Korea

This study aimed to explore variations within native-English-speaking-teachers (NESTs) when evaluating English learners' speaking in terms of their teaching period in one institute (new and experienced) and to compare the results with their writing evaluation behaviors. Using the same methodology to the writing study, the speaking materials were carefully prepared and the speaking delivery performance like pronunciation were excluded to be compatible with the writing evaluation. While the experienced NESTs evaluated Korean English learners' essays more severely than the new NESTs in contrast to the previous studies, there was no difference in speaking. When evaluating the learners' English writing, the experienced group demonstrated similarities to non-native-English-speaking-teachers (NNESTs) in contrast to the new NESTs. When evaluating speaking, the new ones were more consistent in regarding content more substantially while the experienced ones showed a different rating behavior from writing. Lastly, unlike the previous studies, but like the writing study in comparison, the NESTs' perceived-difficulty played a more important role in grading than their perceived-importance regarding grammar, content, and vocabulary. The details of the results were elaborated and the discussions with implications were provided. This study also suggests a new approach to examine rater biases in relation to English learners' expressive skills.

Keywords: NESTs and NNESTs, rater biases, speaking evaluation, new raters, experienced raters

Introduction

"The term [native speaker fallacy] was coined…, which stated that the ideal teacher of English is a native speaker" (Maum, 2002, p. 1). This native speaker fallacy raised many questions and encouraged researchers to explore the differences and similarities between native-English-speaking-teachers (NESTs) and non-native-English-speaking-teachers (NNESTs). For example, Medgyes (2001) explored the perceived differences in their teaching behaviors between the two groups while Matsuda and Matsuda (2001) emphasized the possible collaboration of the two groups to compensate each other's weaknesses.

In addition to teaching, their rating behaviors have also been compared since their different rating behaviors rather than their students' poor performances may yield disadvantages to students (Kobayashi, 1992; Lee, 2009; McNamara, 1996; Schaefer, 2008; Shi, 2001; Shin, 2010; Song & Caruso, 1996; Weigle, 1998). However, in a university where only NESTs teach and evaluate their students unlike middle or high schools[1] in Korea, variations within NESTs have been frequently observed. Based on the

[1] In Korean middle or high schools, Korean English teachers (NNESTs) usually lead the classes with co-teachers, NESTs who do not evaluate students' performances or who take charge only of their expressive skills. In contrast, in universities, NESTs evaluate as well as teach their students rather than co-teaching. Also, in some universities, only

observation, it was examined whether their teaching experiences in one institute would cause any difference in evaluating Korean English learners' essays (Lee, 2016a). Interestingly, that study showed that those with more than five years of teaching experiences in Korea gave the highest scores to grammatically accurate essays, which was reported as NNESTs' rating behaviors rather than NESTs' in the previous studies (Brown, 1991; Lee, 2009), while the new ones gave the highest scores to the well-organized essays.

Based on this interesting observation that the experienced NESTs in Korea rated in a more similar way to the Korean NNESTs than the new NESTs, a series of questions were raised; then, how will they evaluate Korean English learners' speaking? Will they be different from new NESTs as they were for writing? In other words, will experienced NESTs evaluate Korean English learners' speaking in a similar way to their writing? If so, will experienced NESTs evaluate it differently from new NESTs as did they for writing? To answer the questions, this study aimed to explore any possible differences between new and experienced NESTs when evaluating Korean English learners' speaking, adopting the methodology of the previous study about the variations among NEST raters of writing (Lee, 2016a).

Literature Review

Rater differences among NESTs as well as between NESTs and NNESTs have been explored in different contexts including Korea (Brown, 1991; Lee, 2009; Lee, 2016a, 2016b; McNamara, 1996; Shin, 2010; Weigle, 1998; Winke, Gass, & Myford, 2012). Some researchers explored relationships between holistic scores and rating criteria (Shi, 2001) since analytic rubrics with several rating criteria would help raters more systematically than holistic rubrics and subjective holistic scoring (Hamp-Lyons, 1991; Kohn, 2006).

In addition to the benefit of using a rubric for increasing inter-rater reliability (Kohn, 2006), the effectiveness of a rubric has been reported in various aspects, such as improving raters' self-confidence on their evaluation (Stevens & Levi, 2005) and helping raters provide consistent feedback (Spandel, 2006). Accordingly, new raters with the rubric, different from those without it, evaluated English learners' essays as consistent as did experienced raters (Lee, 2016b).

However, even with a rubric, different rating behaviors have been observed within each group (within NESTs and within NNESTs) and between the groups (NESTs and NNESTs). For example, NESTs evaluated the content and organization more severely than they did for the language use and mechanics of Japanese English learners' essays (Schaefer, 2008), while NNESTs evaluated the grammar more severely than they did for the organization of Korean English learners' essays (Shin, 2010). NESTs put less emphasis on grammatical accuracy than NNESTs (Brown, 1991; Lee, 2009), and the former put more emphasis on logical content of an essay than the latter who emphasized its grammatical accuracy (Song & Caruso, 1996).

Most of the studies on rater differences explored variations among raters when evaluating English learners' essays as mentioned above (Cho, 2008). In English learning contexts as in Korea, NESTs are responsible for teaching and evaluating expressive skills; their variations when evaluating learners' speaking should also be examined. Interestingly, in terms of English learners' speaking, raters' rating behaviors have been little explored compared to learners' speaking delivery performance like pronunciation, stress, intonation, speed, volume, and so on (Bijani & Khabiri, 2017; Iwashita, Brown, McNamara, & O'Hagen, 2008). Accordingly, organization (content), grammaticality, and vocabulary of the learners' speaking have been rarely paid attention to even though these are also essential for expressive skills like speaking as well as writing. While speaking rubrics like TOEFL speaking include all (language use for grammar and vocabulary, topic development for content, and delivery for speaking delivery performance), most speaking raters tend to focus heavily on speaking delivery performance

NESTs teach the entire group of freshmen while in some other universities, NESTs teach a part of freshmen.

especially when evaluating English learners' speaking (Iwashita et al., 2008). Even though speaking delivery performance is important for communication, it should be noted that English learners might be misunderstood as poor speakers and might not have a chance to address their thoughts; moreover, raters might not also have a chance to listen to English learners' message simply because of their poor pronunciation. Few studies have explored English learners' expressive skills, especially when speaking, regarding how logically content is organized and how good is their language use. Moreover, variations among raters within NESTs who take most charge of evaluating speaking in Korea have been rarely explored as much as variations among writing raters even though it is highly possible for experienced NESTs to get used to Koreans' specific delivery performance, resulting in different rating behaviors from new NESTs.

Therefore, this study aims to explore whether NESTs with teaching experiences in one institute in Korea more than five years (experienced NESTs) would evaluate English learners' speaking differently from those without any teaching experiences (new NESTs), excluding learners' speaking delivery performance which might prevent raters from concentrating on learners' content, grammar, and vocabulary use. This study examines the following research questions in order to see whether NESTs evaluate Korean English learners' speaking in a similar way to their writing:

(1) Do experienced NESTs evaluate English learners' speaking differently from new NESTs when deciding holistic scores? Is it different from the results of the writing evaluation?
(2) Do experienced NESTs evaluate English learners' speaking differently from new NESTs in terms of grammar, content, and vocabulary? Which criterion are they more sensitive to when rating learners' speaking? Is it different from the results of the writing evaluation?
(3) What criterion do NESTs consider most difficult and important when deciding holistic scores? Is it different from the results of the writing evaluation?

Methods

Participants

The NESTs who have participated in writing evaluation at one university in Korea were informed of this study on comparison between experienced and new NESTs in terms of speaking evaluation. They agreed to participate in this study as well. Eight NESTs participated: four experienced NESTs with more than five years of teaching at the university and four new NESTs. They were compared in terms of the criteria (grammar, content, and vocabulary) in addition to the holistic scores as were they for writing. It was also compared how important and how difficult they considered when evaluating Korean English learners' speaking.

Instrument

Rubric

As mentioned above, speaking delivery performance like pronunciation, stress, intonation, speed, volume, and so on were intentionally excluded in the rubric so that the NESTs could focus on a learner's use of grammar and words and logical organization of the content (Appendix A). The NESTs taught and rated their students' speech with the rubric, which consists of five levels (poor, below average, average, above average, and excellent). The rubric, with 15 as the highest possible score, offers descriptions of grammar, content (including organization and logic of a speech), and vocabulary to determine holistic scores. For example, when a student's speaking is almost always grammatically accurate, consists of a broad range of grammatical structures, is generally coherent, displays logical, clear organization, fully

addresses the task, and/or uses broad, sophisticated vocabulary, then a NEST can give a score of 13, 14, or 15, which would put a student into the "Excellent" category.

Speaking materials to be evaluated

With the very similar rubric to writing, which is very familiar to the NESTs since they had been teaching freshmen how to speak academically and sophomores how to write academic essays in the university, they were asked to rate a student's different version of English speaking[2]. Since this study explores whether experienced NESTs evaluate English learners' speaking differently from new NESTs in terms of grammar, content, and vocabulary as they did for English learners' essays, the speaking materials containing each factor were needed as were the essays; grammatically correct speaking should be compared with grammatically incorrect one; logically organized speaking, with illogically organized one; and speaking with rich vocabulary use, with one with limited vocabulary use.

A model speech transcribed by an NES textbook writer was modified into seven different types as one model speech was modified into seven different types of speech. The topic of the model speech was "Which one do you prefer, online courses or traditional face-to-face classes at your university?" One or more of the features (grammar, content, and organization) was edited while controlling other features to keep the same except for the difference in attention. For example, the speaking material with repeatedly used synonyms and less sophisticated words has the same coherent content with grammatically accurate sentences to the model speech: it used "at any time they want" instead of "at their own convenience" (Some of my friends like online courses because they can take the class at any time they want) and "are not better than" instead of "outweigh" (these good things are not better than any better things of traditional face-to-face classes).

However, unlike the essays, raters might be influenced by a student's accent, speed, intonation, and so on when evaluating; in order to exclude a student's speaking delivery performance, one senior student majoring in English at the university recorded all the eight versions of speech. When recording, she kept the speed, intonation, stress, volume, and so on the same through the eight versions, so that the NESTs were able to focus on the three criteria in attention to decide the holistic scores.

Questionnaires

Right after rating the eight different versions of the student's speaking, the eight NESTs were asked to answer what component, among grammar, content, and vocabulary, was the most important and difficult when deciding a holistic score for each (Lee, 2009). They were also asked to write down their own opinion on what they considered most important and difficult component when evaluating speaking. The questionnaire included questions about whether they have any relationships with Koreans.

Data Collection Procedures

A model speech transcribed by an NES author was extracted from a course-book. It (GCV) was modified into seven different versions, tampering with one or more features similar to the study done by Lev-Ari and Keysar (2012)[3]. As a result, the eight types of speaking materials (one intact and seven modified) were very similar to each other with similar readability scores (Flesch Reading Ease)[4] of the

[2] It should be noted that only one type of speaking, one-way speech instead of two-way conversation, was evaluated since this study aims to compare NESTs' rating behaviors when evaluating English learners' speaking with their essay rating behaviors.

[3] They examined what makes listeners less attentive to NSs compared to NNSs. They changed 11 words of the same text NSs were supposed to say, to control the other variations between NSs and NNSs' stories in order to prepare a text for NNSs.

[4] 0 to 30 (highly difficult to read), 60 to 70 (medium level of difficulty), and 90 to 100 (easy to read).

transcribed texts[5] (46.48 to 59.52) but with a distinct feature for each type (Appendix B). The eight types of speaking materials are defined as: correct grammar, logical content, and rich vocabulary (GCV); correct grammar, coherent content, and limited vocabulary (GCV-); correct grammar, incoherent content, and rich vocabulary (GC-V); correct grammar, incoherent content, and limited vocabulary (GC-V-); incorrect grammar, coherent content, and rich vocabulary (G-CV); incorrect grammar, coherent content, and limited vocabulary (G-CV-); incorrect grammar, incoherent content, and rich vocabulary (G-C-V); and incorrect grammar, incoherent content, and limited vocabulary (G-C-V-).

In other words, all the other features of the eight types of speaking materials were controlled so that the difference can only outstand. For example, the GCV- type has the same coherent content with grammatically correct sentences to the GCV aside from repeatedly using same words and less sophisticated words. A rater was assumed to give the highest score to the first type of speaking material (GCV) and the lowest score to the last type (G-C-V-).

After recording one English learner's speaking with the eight types, maintaining the same speaking delivery performance, the same NESTs were informed of this study. They visited a classroom and listened to the recorded voice by wearing ear-phones to rate the eight types of speaking. Then, they ranked the three factors in the order of difficulty and importance when rating and completed the background questionnaire.

Data Analysis Procedures

Repeated-measure analyses were used to examine how each group rated the eight types of speaking and also to compare whether the two groups rated differently in terms of grammar (G/G-), content (C/C-), and vocabulary (V/V-). If necessary, *t*-tests were also conducted as follow-up tests with an adjusted *p*-value in accordance with the number of pairs to be compared. The results of the previous study on writing (Lee, 2016a) were cited to be compared with those of this study on speaking.

Results

Differences in General

As Table 1 shows, the scores of the two groups were not significantly different, which is different from the writing evaluation. While the new group gave higher scores to the learners' essays than the experienced group in the previous study (Lee, 2016a)[6], there was no difference in terms of speaking evaluation. What might have caused this change, whether the new group became more rigorous or the old group became more lenient to the learners' speaking, was further explored in terms of criteria.

TABLE 1
Differences in General

Group (No. of NESTs)	No. of Speeches	*Mean*	*SD*	*t*	*df*	*p*
Experienced (4)	8	10.38	.95	1.77	6	.127
New (4)	8	9.03	1.18			

[5] GCV (49.28), GCV- (57.12), GC-V (47.37), GC-V- (59.52), G-CV (49.09), G-CV- (58.47), G-C-V (46.48), G-C-V- (57.79)

[6] Lee (2016a) compared the writing scores between the new and the experienced NESTs with much more essays written by all the sophomores in the university. The new group gave significantly higher scores (pre-test 7.43 and post-test 9.04) than the experienced group (pre-test 6.73 and post-test 7.81) to the learners' essays.

The Journal of Asia TEFL
Vol. 15, No. 4, Winter 2018, 1036-1050

Differences in Terms of Criteria

First of all, it was compared whether the NESTs evaluated the eight different types of speaking in the similar way to their writing evaluation. Their mean scores of speaking were compared with those of writing from the previous study in order to see whether they evaluated the eight types of speaking as were expected, the model speaking (GCV) with the highest score to the eighth speaking (G-C-V-) with the lowest score. As presented in Figure 1, the results of the writing evaluation (Lee 2016a) showed the expected hierarchy of the scores even though there was a little variation in the middle, but the results of the speaking evaluation showed quite a big variation in the middle.

Then, in order to see how the new and the experienced groups evaluated the eight different types of speaking, the two groups' scores were compared (Table 2). The previous study showed that the experienced group showed the very similar hierarchy of the scores to the expectation while the new group showed their preference to the coherent and logical content. However, in terms of speaking, as Figure 2 shows, neither group showed the expected hierarchy with difference from each other.

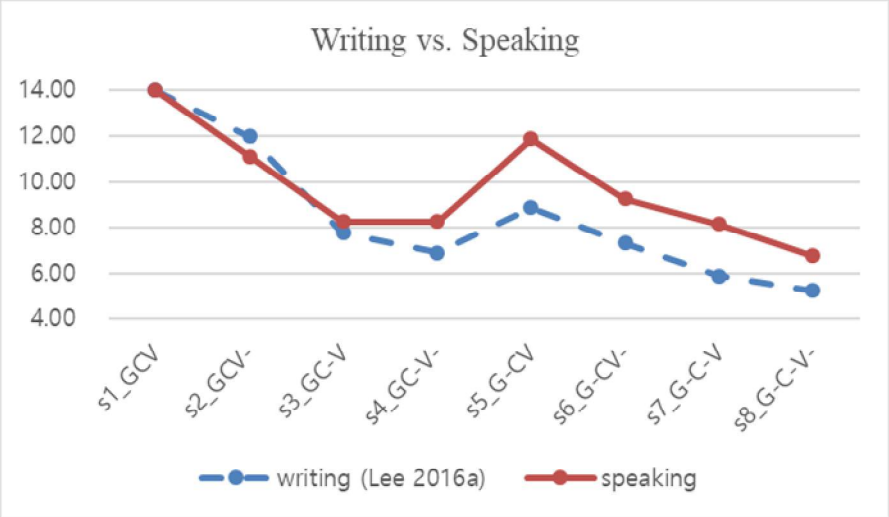

Figure 1. Comparison between writing and speaking evaluation.

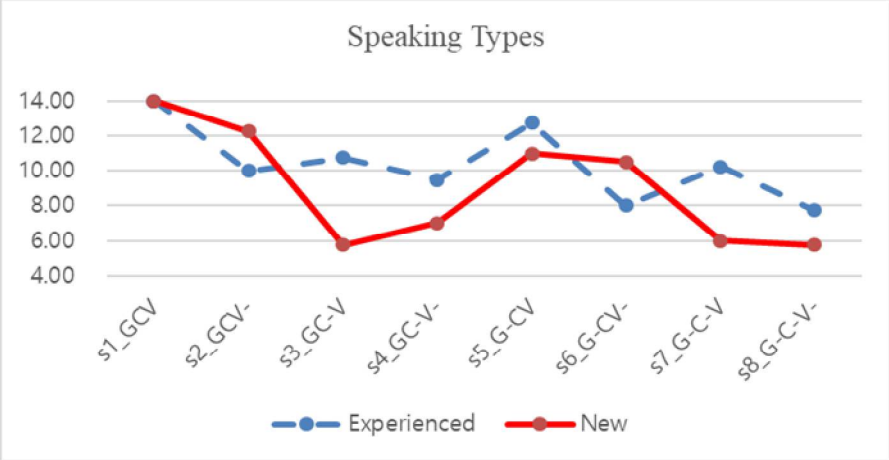

Figure 2. Comparison between new and experienced NESTs in terms of speaking.

In order to see whether the two groups rated differently, a repeated-measure analysis was conducted. As Table 3 shows, the within-subject effect turned out significant (F=15.57, df=7, p=.000, *Effect-Size*=.722); this means the NESTs, regardless of the groups they belonged to, gave different scores to the different types of speaking. The interaction effect between the group and the eight speeches was also significant (F=5.07, df=1, p=.000, *Effect-Size*=.458). The between-subject effect was not significant (F=3.13, df=1, p=.127, *Effect-Size*=.343), but because of the significant interaction effect, the marginal means of the experienced (M=10.38, SD=.95) and the new (M=9.03, SD=1.18) groups were compared using an independent t-test.

As shown in Table 1, the marginal means were not significantly different; however, the follow-up t-tests were conducted with the adjusted p-value (.006=.05/8) to see whether the two groups rated differently any of the eight types of speaking. The grammatical speaking with incoherent content and sophisticated vocabulary use (GC-V) showed a marginally significant difference (t=3.74, df=6, p.=.010) as well as the ungrammatical speaking with coherent content and limited vocabulary use (G-CV-; t=-3.27, df=6, p=.017). In other words, the experienced group was stricter about vocabulary use (the experienced group's mean=8.00 < the new group's mean=10.50) while the new group was stricter about content (the experienced group's mean=10.75 > the new group's mean=5.75).

TABLE 2
Descriptive Statistics

Essay (M by all)	Group (N)	*Mean*	*SD*
GCV	Experienced (4)	14.00	.00
(14.00)	New (4)	14.00	1.15
GCV-	Experienced (4)	10.00	1.41
(11.13)	New (4)	12.25	2.22
GC-V	Experienced (4)	10.75	1.71
(8.25)	New (4)	5.75	2.06
GC-V-	Experienced (4)	9.50	1.91
(8.25)	New (4)	7.00	2.16
G-CV	Experienced (4)	12.75	.96
(11.88)	New (4)	11.00	4.32
G-CV-	Experienced (4)	8.00	.82
(9.25)	New (4)	10.50	1.29
G-C-V	Experienced (4)	10.25	2.22
(8.13)	New (4)	6.00	1.83
G-C-V-	Experienced (4)	7.75	1.71
(6.75)	New (4)	5.75	1.89

TABLE 3
Within-subject Effect (sphericity assumed)

		df		*F*	*P*	*Effect-Size*
Eight Types	326.73	7	46.68	15.57	.000	.722
Eight Types * Group	106.48	7	15.21	5.07	.000	.458
Error	125.91	42	3.00			

The Journal of Asia TEFL
Vol. 15, No. 4, Winter 2018, 1036-1050

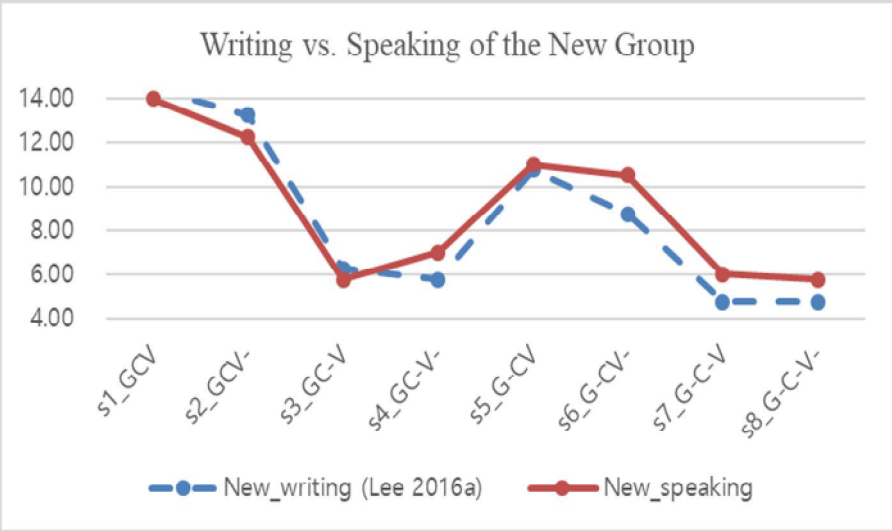

Figure 3. Comparison between writing and speaking scores of the new group.

When comparing the writing scores with the speaking scores respectively given by each of the new and experienced groups, the interesting contrasts were found. As Figure 3 shows, the new group showed very similar results of speaking evaluation to writing, which means that the new NESTs consistently showed the preference to the coherent and logical content regardless of the delivery means. In other words, those who started teaching English in Korea showed consistent rating behaviors to both writing and speaking. In contrast, the experienced group gave the highest scores to the model essay and lower scores to the essays with incorrect grammar, incoherent content, and/or limited use of vocabulary as expected, but not in the same way to speaking (Figure 4). That is, the experienced NESTs did not rated the learners' writing and speaking in a consistent way, as the new NESTs did.

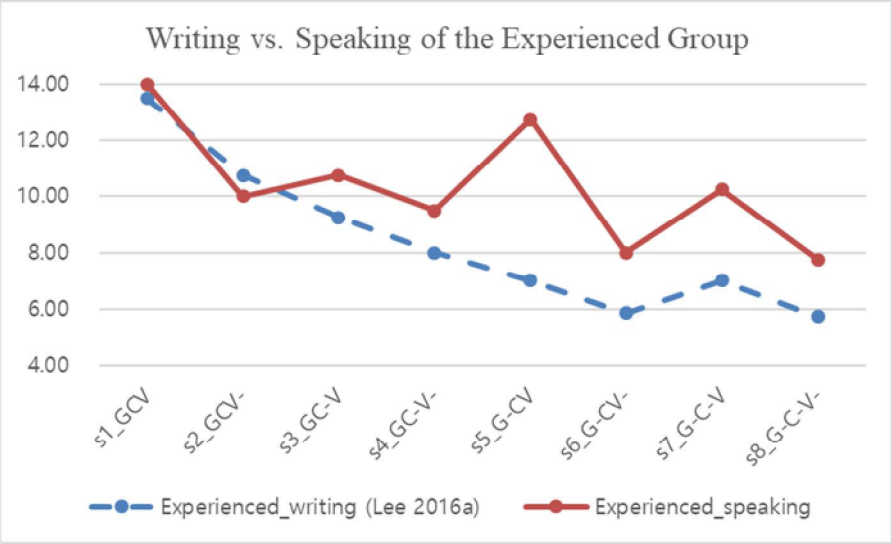

Figure 4. Comparison between writing and speaking scores of the experienced group.

Then, in terms of three criteria, it was again examined whether the good speeches (G/C/V) got higher scores than the poor ones (G-/C-/V-) as were the cases of the writing. As Table 5 shows, the within-subject effect turned out significant (F=27.02, df=5, p=.000, *Effect-Size*=.818); this means the NESTs, regardless of the groups they belonged to, gave different scores to the speaking in terms of six

classification (G/G-/C/C-/V/V-, see Table 4). The interaction effect between the group and the six classification was also significant (F=9.16, df=5, p=.000, *Effect-Size*=.604). The between-subject effect was also significant (F=653.26, df=1, p=.000, *Effect-Size*=.991), so the follow-up t-tests were conducted with the adjusted p-value (.008=.05/6) to see whether the two groups rated differently any of the six classifications of speaking. Marginally significant differences between the two groups were found in terms of good vocabulary use (V; t=2.99, df=6, p=.024) and poor content (C-; t=2.67, df=6, p=.037). In other words, the experienced group preferred the better vocabulary use (the experienced group's mean=11.94 > the new group's mean=9.19) while the new group was stricter about incoherent content (the experienced group's mean=9.56 > the new group's mean=6.13).

TABLE 4
Comparisons in Terms of Criteria

	Group (N)	Good Speech		Poor Speech	
		Mean	*SD*	*Mean*	*SD*
Grammar	Experienced (4)	11.06	.94	9.69	1.16
	New (4)	9.75	.54	8.31	1.88
	Total (8)	10.41	1.00	9.00	1.62
Content	Experienced (4)	11.19	.66	9.56	1.48
	New (4)	11.94	1.21	6.13	1.76
	Total (8)	11.56	.99	7.84	2.38
Vocabulary	Experienced (4)	11.94	1.03	8.81	1.23
	New (4)	9.19	1.78	8.89	.63
	Total (8)	10.56	1.99	8.84	.91

TABLE 5
Within-subject Effect (sphericity assumed)

		df		*F*	*P*	*Effect-Size*
Six Classification	75.04	5	15.01	27.02	.000	.818
Six Classification*Group	25.45	5	5.09	9.16	.000	.604
Error	16.66	30	.56			

As described above with the Figures 3 and 4, Figure 5 shows the new group's more similar rating behaviors to both writing and speaking in terms of three criteria (six classifications) than the experienced group.

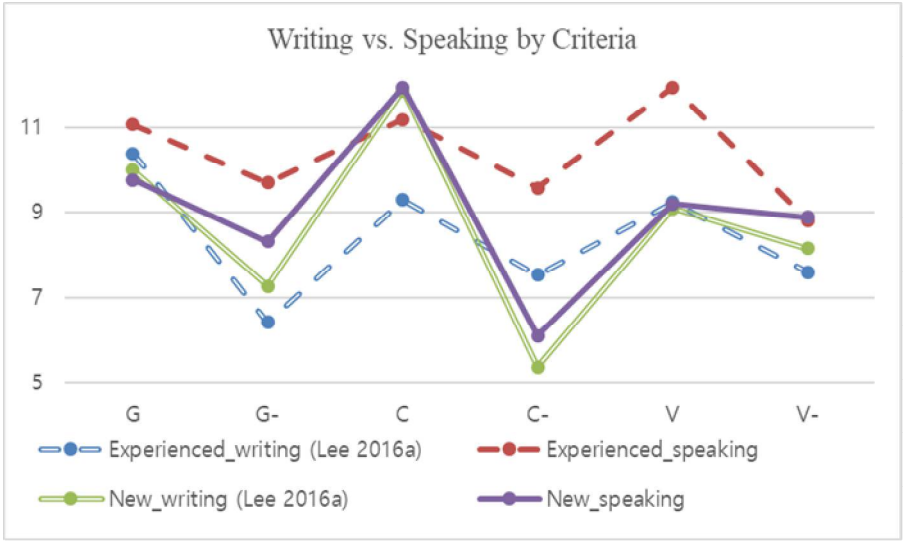

Figure 5. Comparison between writing and speaking scores in terms of criteria.

Based on Figure 5, in terms of criteria, it seems that the experienced NESTs gave higher scores to the good English (speaking and writing) than the poor English (speaking and writing) and that the new NESTs were more affected by content than grammar or vocabulary. To see whether they felt the same way when actually deciding the scores, the raters were asked to give a rank to the three criteria to show their perceived-importance and perceived-difficulty. Since there were not many raters, how many of them ranked each factor as the first one was counted. As Table 6 shows, the experienced NESTs' perceived-importance was not different from the new ones; regarding their perceived-difficulty, no experienced one considered vocabulary most difficult and no new one considered content most difficult. Even though the sample size is small, the distributions regarding speaking looks different from those regarding writing. In other words, the NESTs seem to consider different criteria important and difficult when evaluating learners' writing and speaking.

TABLE 6
Perceived-Importance and Difficulty

Group	Perceived-Importance (Writing; Lee, 2016a)			Perceived-Difficulty (Writing; Lee, 2016a)		
	G	C	V	G	C	V
Experienced	1(1)	2(3)	1(0)	1(4)	3(0)	**0**(0)
New	1(0)	1(1)	2(3)	2(0)	**0**(2)	2(2)

Discussion

The first research question was whether the experienced NESTs would evaluate English learners' speaking differently from the new NESTs when deciding holistic scores. The answer is no. However, the results were different from those of writing evaluation; while the new ones gave higher scores to the learners' essays than the experienced ones in contrast to the previous studies (e.g., McNamara, 1996; Weigle, 1998), there was no group difference in their speaking scores. This may be related to the small sample size. Writing scores of 20 raters (8 experienced and 12 new NESTs in one university) were compared with more than 2,000 essays, but for the speaking, a small number of the raters' (4 experienced and 4 new NESTs) scores about eight speeches were compared.

However, since this study adopted the similar processes to the writing study (Lee, 2016a) to understand the raters' behaviors when evaluating speaking compared to writing evaluation, it could give a meaningful insight to the comparison in focus. Based on the result, it seems that NESTs behaved differently when evaluating English learners' essays from evaluating their speaking performance. Future studies with more raters and more speeches are needed to see whether the other NESTs would behave in a similar way to those in this study regarding speaking and/or writing in terms of their teaching experiences.

Regarding the second research question, "Do experienced NESTs evaluate English learners' speaking differently from new NESTs in terms of grammar, content, and vocabulary?" yes, they did. Considering the results of the eight different types of speeches, even though neither group did not show the expected hierarchy (Figure 2) as the experienced group's writing scores (Figure 4) and all the raters' writing scores (Figure 1) did. With marginal significance, the experienced group tended to be stricter about vocabulary use (G-CV-) while the new group, stricter about content (GC-V). Similarly, considering each criterion in focus, marginally significant differences between the two groups were found in terms of good vocabulary use (V) and poor content (C-); the experienced group preferred the better vocabulary use while the new group was stricter about incoherent content.

This means that the new NESTs regarded content more substantially when evaluating learners' writing and speaking performances, which supports the previous studies (e.g., Shi, 2001; Song & Caruso, 1996) about the NEST raters compared with their NNEST counterparts; however, the experienced NESTs in this study put more emphasis on vocabulary when evaluating speaking while putting more emphasis on grammar when evaluating writing, which is in contrast to the previous studies reporting that NNESTs are

more concerned about grammar than NESTs (Lee, 2009; Shi, 2001; Shin, 2010). This study showed another contrasting result to the other comparative studies showing that experienced raters tend to rate in a relatively more consistent and lenient way than the new ones (McNamara, 1996; Weigle, 1998); the new NESTs in this study were more consistent in valuing content when evaluating English learners' speaking as well as writing than the experienced ones.

However, is consistency good when it is related to different expressive skills, writing and speaking? When the previous studies stated the consistency, it was about rating one kind of skill, mostly writing; few studies have compared the consistency of the same raters' writing evaluation with their speaking evaluation. The results of this study suggest that the experienced raters, who have been developing the deeper understanding of Koreans with more than five years of teaching and staying in Korea, were able to choose different criteria to properly assess the Korean learners' speaking and writing performances. In other words, when Korean English learners write English essays, they may put much energy and time to choose sophisticated words to organize coherent and logical content, resulting in relatively more grammatical mistakes than expected, so the experienced NEST raters, like Korean English teachers, tend to check their grammar mistakes strictly. In contrast, when they speak in English, they may resort to the limited vocabulary in their working memory, resulting in repeating the same words, so the experienced NEST raters tend to check their vocabulary use more strictly. Unlike the experienced NESTs, the new NESTs could not yet figure out these differences with fewer opportunities to adjust their criteria to properly evaluate each skill, resulting in checking the content in the same way. Therefore, English learners' improvement should be examined longitudinally regarding whether the consistency between writing and speaking evaluation would be helpful for improving their performances. In addition, it is needed to study further whether Korean English teachers would put an emphasis on vocabulary when evaluating speaking as much as did the experienced NESTs in this study; this will inform us of the possibility of the experienced NESTs' assimilation (Cherubini, 2009) as Lee (2016a) insisted.

The results of the third research question regarding the criteria seem to be related to the raters' perceived-difficulty. It is possible that both groups used a criterion they felt relatively easy to rely on; no new ones considered content most difficult and no experienced ones considered vocabulary most difficult. Since there is not so much time to examine a test taker's speech as her writing, it seems that the raters used the criterion they felt easy to use. This supports the Lee's (2016a) new perspective that raters' perceived-difficulty would play a more crucial role than their perceived-importance which was assumed to be important (Eckes, 2012; Lee, 2009). However, it should be noted that for the writing evaluation, the criteria each group considered most difficult played an essential role, whereas for the speaking evaluation, the criteria each group considered most easy played an essential role. Based on the results, it is necessary to find out whether the delivery means, speaking or writing, makes this difference and which one, using an easy criterion or a difficult one to decide the holistic scores, would be effective for raters. Also, the raters, regardless of their teaching experiences, should be provided enough opportunities to be aware and understand what criterion they feel difficult or easy, how they can make it easy or difficult, what else they can use to evaluate English learners' performances properly. Regular workshops can be useful that the new and experienced raters share their experiences and feelings and the experienced ones can mentor the new ones to develop good understanding to Korean English learners.

As mentioned above, this study did not explore how the NEST raters evaluate Korean English learners' conversation, which may be more common than the speech format speaking in this study. There are limitations caused by this limited format of speaking, so the future studies comparing the NESTs' rating behaviors in terms of their teaching experiences in Korea should be conducted using other types of speaking.

Conclusion

This study aimed to explore whether the NESTs in one institute in Korea would rate Korean English

learners' speaking in a similar way to their writing in terms of their teaching experiences (new vs. experienced raters). Most studies have been focusing on either one skill, writing or speaking (mostly on writing though), so this study adds a new perspective to the field of rater-basis research. Even though the speaking materials in this study might not be used for everyday conversation, this study adopted the same methods to the previous writing evaluation study to explore the NEST raters' speaking evaluation behaviors. Since the learners' speaking delivery performance like pronunciation, stress, intonation, speed, volume, and so on, which have been playing a major role in speaking evaluation, was intentionally excluded, this study might suggest a new advice to English learners that they should focus on developing the skills how to organize the content coherently and logically with good vocabulary in addition to their speaking delivery performance skills. In addition, the raters would be benefited from the results of this study regarding their perceived-difficulty, and their realization that they might have been using different criterion to evaluate speaking from writing would ultimately benefit their students, English learners.

The Author

Kyoung Rang Lee is an associate professor at Sejong University, Seoul, Korea. She is interested in individual differences in teaching and learning English, including strategies of both teachers and students. Currently, she is devoted to better understanding assessment strategies of native English speaking teachers in Korea as well as promoting Koreans' L2 learning strategy awareness and use.

Department of English Language and Literature,
Sejong University
Gunjadong 98 Gwangjingu, Seoul, 143-747, Korea
Tel: +82 234083118
Email: kranglee@sejong.ac.kr

References

Bijani, H., & Khabiri, M. (2017). Investigating the effect of training on raters' bias toward test takers in oral proficiency assessment: A FACET analysis. *The Journal of Asia TEFL, 14*(4), 687-702.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*, 587-603.

Cherubini, L. (2009). Reconciling the tensions of new teachers' socialization into school culture: A review of the research. *Issues in Educational Research, 19*(2), 83-99.

Cho, D. (2008). Investigating EFL writing assessment in a classroom setting: Features of composition and rater behaviors. *The Journal of Asia TEFL, 5*(4), 49-84.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly,* 9, 270-292.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5-15). Westport, CT: Ablex Publishing.

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

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*, 81-112.

Kohn, A. (2006). The trouble with rubrics. *English Journal, 95*(4), 12-15.

Lee, H. (2009). Native and nonnative rater behavior in grading Korean students' English essays. *Asia Pacific Education Review, 10*(3), 387-397.

Lee, K. (2016a). Diversity among NEST raters: How do new and experienced NESTs evaluate Korean English learners' essays? *The Asia-Pacific Education Researcher, 25*(4), 549-558.

Lee, K. (2016b). The effects of a rubric on inexperienced raters' scoring consistency. *Modern English*

Education, 17(2), 75-90.

Lev-Ari, S., & Keysar, B. (2012). Less-detailed representation of non-native language: Why non-native speakers' stories seem more vague. *Discourse Process, 49*(7), 523-538.

Matsuda, A., & Matsuda, P. K. (2001). Autonomy and collaboration in teacher education: Journal sharing among native and nonnative English-speaking teachers. *CATESOL Journal, 13*(1), 109-121.

Maum, R. (2002). Nonnative-English-speaking teachers in the English teaching profession. Eric Digest. EDO-FL02-09. Retrieved 18 October 2017 from http:www.cal.org/resources/digest/0209 maum.html

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

Medgyes, P. (2001). When the teacher is a non-native speaker. *Teaching Pronunciation*, 429-442. Retrieved 18 October 2017 from http://teachingpronunciation.pbworks.com/f/When+the+teacher+ is+a+ non-native+speaker.PDF

Schaefer, E. (2008). Rater bias patterns in EFL writing assessment. *Language Testing, 25*(4), 465-493.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

Shin, Y. (2010). A FACETS analysis of rater characteristics and rater bias in measuring L2 writing performance. *English Language & Literature Teaching, 16*(1), 123-142.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing,* 5(2), 163-182.

Spandel, V. (2006). In defense of rubrics. *English Journal, 96*(1), 19-22.

Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback and promote student learning.* Sterling, VA: Stylus.

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

Winke, P., Gass, S., & Myford, C. M. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*, 231-252.

Appendix A

Speaking Assessment Rubric

Assessment	Score	Response
Excellent	13, 14, 15	-Grammatically accurate; contains some minor errors; displays a broad range of grammatical structures; displays broad, sophisticated vocabulary -Completely coherent and well organized and exhibits excellent use of cohesive devices -Is completely intelligible -Speaker's meaning is clearly conveyed and ideas mentioned are fully developed -Fully addresses the task
Above Average	10, 11, 12	-Mostly grammatically accurate; may contain some minor errors and a few major errors; displays a relatively broad range of grammatical structures; displays a relatively wide range of vocabulary -Is generally coherent and well organized; displays good use of cohesive devices -Is generally intelligible -Speaker's meaning is generally clear and the ideas are fairly well developed -Adequately addresses the task
Average	7, 8, 9	-Fairly accurate; may contain occasional major errors; is somewhat limited to simple sentences; displays somewhat narrow vocabulary -Is at times incoherent and may contain parts that display unclear organization; displays use of simple cohesive devices (e.g. repetition of *and* or *but*) -Is sometimes unintelligible -Speaker's meaning is at times obscure and the ideas mentioned could be more developed -Adequately addresses the task
Below Average	4, 5, 6	-Contains several major and minor errors; is limited to a narrow range of grammatical structures; is often limited to simple sentences; displays a simple and limited vocabulary -Is often incoherent and rather loosely organized; displays use of simple cohesive devices that aren't always effective -Is often unintelligible -Speaker's meaning is often unclear and the ideas mentioned are rather inadequately developed -Marginally addresses the task
Poor	1, 2, 3	-Almost always grammatically inaccurate; displays only simple sentences and extremely limited and simple vocabulary -Is generally incoherent; displays illogical, unclear organization -Is generally unintelligible -Speaker's meaning is generally unclear -Barely addresses the task

Appendix B

Readability of the Transcribed Speaking Materials

Type	No. of Words	No. of Sentences	Flesch Reading Ease	Feature
GCV	20.40	10.00	49.28	A model speech transcribed by NES author
GCV-	17.80	10.00	57.12	Synonyms and less sophisticated words were repeatedly used.
GC-V	19.36	11.00	47.37	Incoherent statements were added and/or the logical flow of the speech was damaged.
GC-V-	15.09	11.00	59.52	Synonyms and less sophisticated words were used and incoherent statements were added and/or the logical flow of the speech was damaged.
G-CV	17.09	11.00	49.09	Grammatically wrong sentences and/or fragments were added.
G-CV-	15.55	11.00	58.47	Grammatically wrong sentences and/or fragments were added and less sophisticated words were used.
G-C-V	17.33	12.00	46.48	Grammatically wrong sentences and/or fragments with incoherent statements were added.
G-C-V-	15.00	11.00	57.79	Grammatically wrong sentences and/or fragments with incoherent statements were added and less sophisticated words were used.