Modern English Education, Vol. 17, No. 2, Summer 2016
http://dx.doi.org/10.18095/meeso.2016.17.2.04

The Effects of a Rubric on Inexperienced Raters' Scoring Consistency

Kyoung Rang Lee
Sejong University

Lee, Kyoung Rang. (2016). The effects of a rubric on inexperienced raters' scoring consistency. *Modern English Education*, *17*(2), 75-90.

This study aimed to explore whether a rubric was effective for promoting raters' scoring consistency when evaluating English learners' essays. The experienced raters' scores of Korean university students, with a rubric, were analyzed qualitatively and quantitatively. The scores of inexperienced raters with the rubric were compared with those without the rubric. Each group's scores were compared using t-tests (also with inter-rater reliability tests) and the inexperienced groups' scores were compared using ANOVA. With the rubric, the inexperienced raters showed a similar way of evaluating students' performance to the experienced raters while those without the rubric could not. The former scored the essays consistently both within and across the semesters, but the latter could do so only within the semesters (post-test scores were higher than pre-test scores within each semester). The scores of each test given by the inexperienced raters with the rubric were not statistically different, which was not true for the experienced with the rubric. This implies that rubrics are effective in promoting inexperienced raters' scoring consistency, especially when raters evaluate essays at the same time rather than when evaluating at different times even with a rubric. The details of the results are given, and the significance is discussed.

**[rubric/scoring consistency/intra-rater reliability/
루브릭/채점일관성/채점자내신뢰도]**

I. INTRODUCTION

A few years ago, a university located in Seoul, Korea tried to administer a writing test to all freshmen, and the English teachers (called raters hereafter to emphasize their role as a rater, even though they were also teaching writing classes) were provided a scoring guide with rating criteria (an analytic rubric) to grade the essays. A rubric refers to scoring guides

consisting of rating criteria, describing how students are supposed to complete a given task and its specifications (Stevens & Levi, 2005). However, the raters resisted to use the rubric because they thought it was unnecessary to check each item repeatedly and sum the scores of each part. They insisted that their professional experiences as English teachers for several years helped them grade English learners' essays holistically, and they did not value the analytic rating criteria (e.g., grammar, content, and vocabulary) of the rubric. Most of them did not major in English education, so the analytic rubric was quite new to them. This let me wonder what a rubric can do for those who have never used it before.

Moreover, the average scores of the essays that each rater gave were different by their environment; in particular, some raters gave more severe scores in the morning whereas others gave more lenient scores after lunch. They also disliked to score two tests (pre-test and post-test) at the end of the semester altogether for examining the students' possible improvement within the semester as they were in a scoring workshop. Without a rubric, not comparing each student's post-test with the pre-test, their scoring did not reflect the freshmen's improvement in writing proficiency even though they taught them two writing classes consecutively for one year; in other words, they ultimately gave scores to their essays without comparing the initial performance at the beginning of the semester, which would have been possible with a rubric.

A rubric helps raters provide their students with consistent feedback (Beyreli & Ari, 2009; Spandel, 2006) based on the objective analytic or holistic criteria rather than based on their subjective holistic scoring.[1] Thus, it promotes the raters' self-confidence on their evaluation[2] (Silvestri & Oescher, 2006; Stevens & Levi, 2005), and moreover, it improves inter-rater reliability (Kohn, 2006; Rezaei & Lovorn, 2010).

In terms of essay raters' inter-rater reliability, a fair amount of research has been conducted to explore what influences different rating behavior of raters (rater bias) when evaluating English learners' writing performance (Brown, 1991; Hyland & Anan, 2006; Kobayashi, 1992; H. Lee, 2009; McNamara, 1996; Schaefer, 2008; Shi, 2001; Y. Shin, 2010; Song & Caruso, 1996; Weigle, 1998) and whether rater-training reduces rater biases (Brown, 1995; Y. Choi, 2002; Knoch, Read, & von Randow, 2007; Kondo-Brown, 2002; McNamara, 1996; Weigle, 1998; Wigglesworth, 1993).

Thus, researchers and educators have started to use caution in relation to rater biases so as not to cause disadvantages to English learners. While inconsistent results have been

---

[1] Hamp-Lyons (1991) differentiated subjective holistic scoring from holistic scoring. The former refers to evaluation based on a rater's subjective judgment or impression on an essay, and the latter refers to evaluation based on holistic rating criteria.

[2] According to Bacha (2001), *evaluation* refers to "assigning a score to a direct writing product (p.372)" in this study.

reported regarding the effectiveness of rater-training (Knoch et al., 2007; Kondo-Brown, 2002), a rubric has shown its usefulness both for raters and students (Beyreli & Ari, 2009; Kohn, 2006; Rezaei & Lovorn, 2010; Spandel, 2006; Stevens & Levi, 2005). However, there are still some concerns regarding a rubric in that it can act like a straitjacket preventing learners from developing their creativeness in learning and in that it is time- and energy-consuming to create a good one (Wolf & Stevens, 2007). Even with these concerns, it is necessary to examine what it can do for raters, especially for inexperienced raters since more than 3,000 pre-service English teachers graduate from universities and about 600 new English teachers are recruited at public schools each year[3] (S. Cho et al., 2008).

II. LITERATURE REVIEW

A rubric is used because it helps raters provide feedback effectively and learners self-evaluate their essays as well (Beyreli & Ari, 2009). It provides consistent rating criteria for open-ended tasks which might be criticized due to lack of objectivity, resulting in promoting raters' self-confidence on evaluation (Silvestri & Oescher, 2006; Stevens & Levi, 2005), and it is reported to improve inter-rater reliability between raters (Kohn 2006; Rezaei & Lovorn, 2010).

Analytic rubrics are generally used to provide scores of sub-skills (such as grammar, vocabulary, and content) as well as a score or a grade (sum of the scores of the elements) of an essay while holistic rubrics are generally used to provide a score or a grade for an essay (Weigle, 2002). When raters evaluate students' open-ended tasks such as English writing based on a holistic rubric, the rubric guides them to evaluate the tasks consistently as well, but without a rubric, raters depend too much on their subjective judgment or impression on an essay (Hamp-Lyons, 1991).

As one example of a holistic rubric, TOEFL writing performance is evaluated according to whether a learner properly selects information to address a given topic and coherently presents the information with less grammar and language errors (Weigle, 2002). An analytic rubric, such as ESL Composition Profile, consists of content, organization, vocabulary, language use, and mechanic (Jacobs, Zingkgraf, Wormuth, Hartfiel, & Hughey, 1981).

Regarding these rating criteria, Shi (2001) compared the two groups' holistic scores as well as five rating criteria (general, content, organization, language, and length) and reported results giving helpful implications; native English speaking raters showed no significant differences from non-native English speaking raters when assigning holistic

[3] They are all English teachers at secondary schools.

scores of their students' essays. However, when looking into the five rating criteria in detail, the latter group was influenced negatively by the organization and length of the essays while the former, positively by the content and language. Brown (1991) compared English faculties with ESL faculties in terms of evaluation focus. Coherence and grammar were considered important by English faculties while organization was focused on by ESL faculties. O'Laughlin (1992) also observed that ESL teachers focused more on organization and content of the essays written by native speakers of English while considering grammar and coherence more on the essays written by English learners.

In addition, raters' scoring consistency in terms of their experience was compared, and inexperienced raters were observed to evaluate inconsistently unlike experienced ones. McNamara (1996) compared experienced raters with inexperienced ones and found out that the former gave higher scores to grammatically correct essays even though their perceptions of importance on grammatical accuracy was not high. Weigle (1998) also compared experienced with inexperienced raters in terms of severity and consistency before and after rater training. The results showed that inexperienced raters tended to evaluate students' essays more severely and less consistently than experienced ones.

As summarized, a rubric itself shows its effectiveness for raters, but when comparing different groups of raters (e.g., native English speaking vs. non-native English speaking raters, English faculties vs. ESL faculties, and experienced vs. inexperienced raters), it has been used differently. Few studies have examined its effectiveness for inexperienced raters with the students' longitudinal writing performance (for one year) whatsoever. In other words, it has not been much examined whether a rubric is helpful for new teachers who have never rated before in Korea and whether it maintains raters' scoring consistency when scoring writing performances at different periods of time as well as at the same time together. Thus, it is necessary to explore whether a rubric does help raters evaluate English learners' essays consistently and reliably tracking their progresses well both within semesters and across semesters. This study examines the following research questions:

1) Do inexperienced raters with a rubric (scoring the essays at the same time together) behave differently from experienced raters with a rubric (scoring the essays right after each test; four different period of time) in terms of scoring consistency?
2) How do inexperienced raters with a rubric evaluate Korean English learners' writing performance for one year? Do they do differently from those without a rubric?

III. METHOD

1. Participants

In order to explore the effectiveness of a writing rubric, an institute administering writing tests regularly to track students' progress was needed. A four-year university in Seoul, Korea was chosen to recruit participants for this study. The university is ideal for this study since two writing tests are administered each semester for one year to all sophomores (about 2,500) in mandatory writing classes. All sophomores are required to take the two consecutive writing classes and the two writing tests per semester. For this study, native English speaking raters were not only the students' primary instructors but also rated all their students' essays with a given rubric.

For the first phase, the scores of Korean English learners'essays were examined in order to see whether the experienced raters with a rubric evaluated learners' improvement especially in regard to their initial writing proficiency. In the first semester, Korean English learners' essays (2,474 in total) were evaluated by 22 raters at the beginning and end of the semester. In the second semester, almost all students took the second writing classes except for those who took or returned from a leave of absence. They also wrote the two essays (2,577 in total) and were evaluated by 22 raters.

Though nearly all students took the writing classes for one year, not all took the four writing tests (pre-test and post-test in the first semester and pre-test and post-test in the second semester). Those who did not take all four tests were excluded, which left 1,157 students' essays evaluated by 12 raters for analysis.

For the second phase, to compare the experienced raters' scoring consistency with the inexperienced ones each of whom scored all the four essays, those who were scored by different raters for two consecutive semesters were removed, thus 580 students' essays evaluated by 8 raters were ultimately analyzed. Then, inexperienced raters were recruited from the pre-service teachers. Eight volunteers decided to score the 580 students' essays

TABLE 1

Profile of the Participants

	No. of Raters	No. of Students	No. of Essays	Description
1st Phase	12	1,157	13,884	The essays of those who took all four tests were analyzed.
2nd Phase	8	580	2,320	The essays of those who were scored by the same raters were analyzed. They rated the essays at different periods of time, right after each test. Inexperienced pre-service teachers with and without a rubric rated the essays. They rated the essays at the same time.

(they rated 2,320 (580 students x 4 essays) essays). Four of them were provided the same rubric used by the raters in the first phase, and the other four were not.

2. Instruments

The raters were provided with the same writing rubric (Appendix 1) to base their teaching and to rate students' essays. The rubric consists of five levels (poor, below average, average, above average, and excellent) with a possible high score of 15. The rubric provided descriptions of grammar, content (includes organization and logic of an essay but will be called 'content' hereafter), and vocabulary in order to determine holistic scores. For example, when an essay is almost always grammatically inaccurate, is generally incoherent, displays only simple sentences and illogical, unclear organization, barely addresses the task, and/or uses extremely limited and simple vocabulary, then a rater gives a score of 1, 2, or 3, which would put the essay into the "Poor" category.

Based on the rubric, raters taught the students how to write academic essays focusing on grammar, content, and vocabulary. The students were required to write an English essay at the beginning and end of the semester to get a grade in the required course.

3. Data Collection Procedures

For the first phase, before the first semester began, a workshop was held for the raters in the university so that they could become normed to the rubric. They were informed that the two writing tests should be administered at the beginning and end of the semester. The raters were given the test sheets and administered the test to their students for about 40 minutes in the second week of the semester. After they graded their own students' essays, both the scores and essays were submitted. Then, raters instructed their students on academic essay composition, concentrating on grammar, content, and vocabulary. In the 14th week of the semester, they repeated the process for a second time, administered the second test, and submitted those results.

As was in the first semester, a workshop was held for the raters with the same rubric in the second semester. They administered the third test in the second week and the fourth test in the 14th week of the second semester.

After collection of data, the scores of the four tests were filtered as described in above section, resulting in 1,157 student participants in total. Two experienced English teachers with more than 10 years of teaching experiences examined each group of the four tests to see whether writing proficiency improved. Their scores were statistically compared.

For the second phase, the 1,157 students' raters were examined. In order to compare the experienced raters' scoring consistency for one year with the inexperienced ones who

scored the four tests at the same time, only those who were scored by the same raters for the both semesters were extracted, resulting in 580 students in total. The same two experienced English teachers looked through each group of the four tests again to see whether their writing proficiency improved as well. Their scores were also statistically compared.

Thereafter, pre-service teachers who had never rated university students' essays were recruited. Eight pre-service teachers volunteered to evaluate the given essays. Four of them were given the rubric and the other four were given a blank sheet of paper without the rubric. Essays were provided in a random order.

4. Data Analysis Procedures

In the first phase, to compare the four writing tests, six dependent t-tests were conducted to check the students' writing proficiency and their improvement (Table 2): 1) post-test to pre-test in the first semester, 2) pre-test in the second semester to post-test in the first semester, 3) post-test to pre-test in the second semester, 4) post-test in the second semester to pre-test in the first semester, 5) pre-test in the second semester to pre-test in the first semester, and 6) post-test in the second semester to post-test in the first semester.

TABLE 2

Reasons of the Comparison

Comparison	Test (semester)	Reasons
1	Post-test (1st) Pre-test (1st)	To see whether students' writing proficiency improved after they learned how to write academic essays in the first semester.
2	Pre-test (2nd) Post-test (1st)	To see whether they maintained their writing ability after the summer vacation.
3	Post-test (2nd) Pre-test (2nd)	To see whether their writing proficiency improved after they learned in the second semester.
4	Post-test (2nd) Pre-test (1st)	To see whether their writing proficiency improved after they learned for one year (two consecutive semesters).
5	Pre-test (2nd) Pre-test (1st)	To see whether their writing proficiency changed before each semester began.
6	Post-test (2nd) Post-test (1st)	To see whether their writing proficiency changed after each semester ended.

In the second phase, the six dependent t-tests both for the scores by the experienced raters in the university and for those by the inexperienced pre-service teachers were conducted. The inter-rater reliability of the experienced raters was checked for each of the four tests (Cronbach's alpha). An analysis of variances (ANOVA) was run as well to see whether the four inexperienced raters of each group rated differently from each other.

IV. RESULTS AND DISCUSSION

1. Experienced Raters

In order to examine whether a rubric helps raters evaluate Korean English learners' writing performance in a consistent way, the scores of the four essays were compared. In other words, it was assumed that a rubric let raters evaluate each essay with the same rating criteria so as to reflect the improvement of the students' writing proficiency without comparing the four essays altogether.

Before conducting the statistical analysis, the two experienced English teachers randomly selected students and compared their four essays. The first essay was written at the beginning of the first semester before they took the required writing course (which was supposed to reflect their initial writing proficiency); the second essay was written after the course was taught for one semester (which was assumed to be improved better than the first, initial writing); the third essay was written at the beginning of the second semester, after summer vacation (which was assumed to be maintained as well as the second essay); and, the fourth essay was written after students learned how to write academic essays for two semesters (which was assumed to be the best).

Even though the two experienced English teachers could not examine all the essays (13,884 essays of 1,157 students), 87 percent (87 out of 100 students) showed the expected improvement in their writing based on the writing criteria. The other 13 students did not show any significant improvement throughout the four essays and used similar grammatical patterns and expressions. Therefore, the rubric was expected to help the raters evaluate the students' essays consistently even though they did not rate them at the same time as the two experienced English teachers did.

Generally speaking, the statistics showed that the experienced raters gave higher scores to the later essays than the first essays as assumed with a little variation. As Table 3 shows, the scores represent the significant improvement of the students' writing proficiency compared to the first essay (7.52 to 9.03, t = 24.44, df = 1,156, p = .000), significant degradation of their performance compared to the second essay (9.03 to 8.11, t = -9.55, df = 1,156, p = .000), and another significant improvement compared to the third essay (8.11 to 9.66, t = 26.27, df = 1,156, p = .000). Also, compared to the initial writing proficiency, the scores showed that their writing significantly improved at the end of the year (7.52 to 9.66, t = 20.92, df = 1,156, p = .000). When comparing their writing before each semester began (7.52 to 8.11, t = 6.03, df = 1,156, p = .000) and after each semester ended (9.03 to 9.66, t = 6.34, df = 1,156, p = .000), both showed significant improvement. It means that the experienced raters were able to reflect the students' writing proficiency in a consistent way as if they rated each student's snapshot of four essays at the same time since they used

the same rating criteria at different rating periods.

TABLE 3

Dependent *t*-Tests (1,157 Students' Essays)

Comparison	Test (semester)	No. of Essays	Mean	*SD*	*t*	*df*	*p*
1	Post-test (1st)	1,157	9.03	2.43	24.44	1156	.000
	Pre-test (1st)	1,157	7.52	2.41			
2	Pre-test (2nd)	1,157	8.11	2.22	-9.55	1156	.000
	Post-test (1st)	1,157	9.03	2.43			
3	Post-test (2nd)	1,157	9.66	2.28	26.27	1156	.000
	Pre-test (2nd)	1,157	8.11	2.22			
4	Post-test (2nd)	1,157	9.66	2.28	20.92	1156	.000
	Pre-test (1st)	1,157	7.52	2.41			
5	Pre-test (2nd)	1,157	8.11	2.22	6.03	1156	.000
	Pre-test (1st)	1,157	7.52	2.41			
6	Post-test (2nd)	1,157	9.66	2.28	6.34	1156	.000
	Post-test (1st)	1,157	9.03	2.43			

2. Inexperienced Raters

Before exploring whether inexperienced raters with a rubric evaluated writing performance differently from those without a rubric, only those who were taught by the same raters for two consecutive semesters were chosen. The scores of the 580 students' essays were compared again to examine their scoring consistency is actually secured throughout the four essays. As Table 4 shows, their improvement was similar to that of the 1,157 students presented as in Table 3.

TABLE 4

Dependent *t*-Tests (580 Students' Essays)

Comparison	Test (semester)	No. of Essays	Mean	*SD*	*t*	*df*	*p*
1	Post-test (1st)	580	9.06	2.25	16.01	579	.000
	Pre-test (1st)	580	7.76	2.28			
2	Pre-test (2nd)	580	8.07	2.08	-8.78	579	.000
	Post-test (1st)	580	9.06	2.25			
3	Post-test (2nd)	580	9.87	2.26	21.05	579	.000
	Pre-test (2nd)	580	8.07	2.08			
4	Post-test (2nd)	580	9.87	2.26	15.87	579	.000
	Pre-test (1st)	580	7.76	2.28			
5	Pre-test (2nd)	580	8.07	2.08	2.64	579	.009
	Pre-test (1st)	580	7.76	2.28			
6	Post-test (2nd)	580	9.87	2.26	6.55	579	.000
	Post-test (1st)	580	9.06	2.25			

These essays were randomly assigned to the two groups of the pre-service teachers, four with and the other four without the rubric. Interestingly, the rubric did help the inexperienced raters (pre-service teachers) evaluate the essays in a consistent way as the experienced raters in the university did for the 580 students (Table 5) even though the inexperienced raters scored a little more severely than the experienced raters.

TABLE 5

Dependent *t*-Tests: Inexperienced Raters With the Rubric

Comparison	Test (semester)	No. of Essays	Mean	*SD*	*t*	*df*	*p*
1	Post-test (1st)	290	8.17	2.19	12.09	289	.000
	Pre-test (1st)	290	6.96	2.15			
2	Pre-test (2nd)	290	7.40	1.75	-4.81	289	.000
	Post-test (1st)	290	8.17	2.19			
3	Post-test (2nd)	290	9.67	2.25	19.95	289	.000
	Pre-test (2nd)	290	7.40	1.75			
4	Post-test (2nd)	290	9.67	2.25	16.02	289	.000
	Pre-test (1st)	290	6.96	2.15			
5	Pre-test (2nd)	290	7.40	1.75	2.75	289	.006
	Pre-test (1st)	290	6.96	2.15			
6	Post-test (2nd)	290	9.67	2.25	9.13	289	.000
	Post-test (1st)	290	8.17	2.19			

TABLE 6

Dependent *t*-Tests: Inexperienced Raters Without the Rubric

Comparison	Test (semester)	No. of Essays	Mean	*SD*	*t*	*df*	*p*
1	Post-test (1st)	290	9.63	1.98	6.62	289	.000
	Pre-test (1st)	290	8.66	2.27			
2	Pre-test (2nd)	290	8.21	1.80	-10.22	289	.000
	Post-test (1st)	290	9.63	1.98			
3	Post-test (2nd)	290	8.49	1.81	2.41	289	.017
	Pre-test (2nd)	290	8.21	1.80			
4	Post-test (2nd)	290	8.49	1.81	-.88	289	.382
	Pre-test (1st)	290	8.66	2.27			
5	Pre-test (2nd)	290	8.21	1.80	-2.84	289	.005
	Pre-test (1st)	290	8.66	2.27			
6	Post-test (2nd)	290	8.49	1.81	-7.91	289	.000
	Post-test (1st)	290	9.63	1.98			

However, without the rubric, it seemed that the raters scored the essays based on their subjective impressions, which resulted in unreliable scores not reflecting the students' improvement (Table 6). Their scores showed no significant improvement from the first essays to the last essays (8.66 to 8.49, t = -.88, df = 289, p = .382) even though there was significant improvement in the first semester (8.66 to 9.63, t = 6.62, df = 289, p = .000) and

the second semester (8.21 to 8.49, t = 2,41, df = 289, p = .017), respectively. Furthermore, their scores in the first semester were higher than those in the second semester (8.66 to 8.21, t = -2.84, df = 289, p = .005 and 9.63 to 8.49, t = .7.91, df = 289, p = .000), which means that the raters without the rubric were not able to evaluate the essays consistently across the semesters although they could within semesters.

In general, the rubric helped the raters, both the experienced and the inexperienced ones, evaluated Korean English learners' essays consistently. However, even though the scores of 1,157 students showed the assumed improvement throughout the semester and the year, more essays should be examined qualitatively than the two English teachers did in this study. Also, qualitative interviews should be conducted to explore what made the inexperienced raters without the rubric gave lower scores to the third and the fourth essays than the first and the second ones while those with the rubric reflected the students improvement consistently. Having two English teachers to review the students' progress qualitatively may be too small of a sample size, so qualitative examination with more reviewers should be conducted.

In order to see whether there were any rater differences within each group, an ANOVA was run. With the rubric, the inexperienced raters' average scores were very similar to each other's on each of the four tests (Table 7), but the scores of those without the rubric were significantly different (pre-test (1st): F = 15.43, df = 3, p = .000; post-test (1st): F = 16.56, df = 3, p = .000; pre-test (2nd): F = 41.73, df = 3, p = .000; and, post-test (2nd): F = 26.97, df = 3, p = .000) as Table 8 shows.

To sum, in terms of comparing the scoring consistency within each group, the inexperienced raters with the rubric showed almost the same results throughout the four essays. It is an unexpected result, compared to the fact that the experienced raters' evaluation for the 580 students significantly differed from each other (pre-test (1st): F = 19.22, df = 7, p = .000; post-test (1st): F = 31.15, df = 7, p = .000; pre-test (2nd): F = 28.55, df = 7, p = .000; and post-test (2nd): F = 35.61, df = 7, p = .000) as the inexperienced raters without the rubric. Moreover, the inter-rater reliability of the experienced raters for each test was not consistently high. The first two tests (pre-test and post-test of the first semester) showed very low inter-rater reliability scores (Cronbach's alphas: .387 and .226), while the latter two tests (pre-test and post-test of the second semester) showed quite high inter-rater reliability scores (Cronbach's alphas: .899 and .683). Similarly, the inter-rater reliability scores of the inexperienced raters without the rubric were not very high throughout the four test: .357 (1st pre-test), .339 (1st post-test), .103 (2nd pre-test), and .424 (2nd post-test).

In contrast, the inter-rater reliability scores of the inexperienced raters with the rubric were quite high throughout the four tests: .689 (1st pre-test), .706 (1st post-test), .310 (2nd pre-test), and .676 (2nd post-test). It seems that the raters' scoring consistency was

promoted when the raters evaluated the essays at the same time as the inexperienced raters did, rather than when they evaluated the essays at different periods of time as the experienced raters did. In other words, the rubric seemed to help the inexperienced raters evaluate learners' essays consistently based on the same rating criteria and reflect both their initial writing proficiency and improvement. Further studies are needed to explore the effects of a rubric when inexperienced raters (and experienced raters as well when possible) evaluate essays at the same time or at different periods of time.

TABLE 7

Differences Within Each Group: Inexperienced Raters With the Rubric

Test	Rater	No. of Essays	Mean	*SD*	*F*	*df*	*p*
Pre-test (1st)	A	83	6.89	2.01	.34	3	.797
	B	66	6.82	2.23			
	C	75	6.97	2.18			
	D	66	6.96	2.18			
Post-test (1st)	A	83	8.13	2.13	.023	3	.995
	B	66	8.23	2.19			
	C	75	8.17	2.07			
	D	66	8.16	2.44			
Pre-test (2nd)	A	83	7.34	1.63	1.41	3	.239
	B	66	7.61	1.82			
	C	75	7.57	1.80			
	D	66	7.06	1.76			
Post-test (2nd)	A	83	9.64	2.27	.67	3	.569
	B	66	9.98	2.38			
	C	75	9.63	2.07			
	D	66	9.44	2.31			

TABLE 8

Differences Within Each Group: Inexperience Raters Without the Rubric

Test	Rater	No. of Essays	Mean	*SD*	*F*	*df*	*p*
Pre-test (1st)	E	91	9.85	1.58	15.43	3	.000
	F	75	8.49	2.27			
	G	62	7.97	2.40			
	H	62	7.79	2.29			
Post-test (1st)	E	91	9.30	1.55	16.56	3	.000
	F	75	10.60	1.99			
	G	62	8.53	1.82			
	H	62	10.06	2.03			
Pre-test (2nd)	E	91	8.85	1.76	41.73	3	.000
	F	75	8.53	1.61			
	G	62	6.32	1.14			
	H	62	8.79	1.31			
Post-test (2nd)	E	91	7.69	.84	26.97	3	.000
	F	75	9.15	1.80			
	G	62	7.73	1.47			
	H	62	9.65	2.23			

As the previous studies showed (McNamara, 1996; Weigle, 1998), the inexperienced raters gave more severe scores than the experienced raters, even with the same rubric. The effectiveness of the rubric in terms of inter-rater reliability seems doubtable again. Thus, in order to improve inter-rater reliability between raters, as to not possibly disadvantage students, trials other than using a rubric, such as evaluating-workshops when raters evaluate essays at the same time together or showing sample essays of each score to refer to, are needed.

V. CONCLUSION

This study aimed to explore whether a rubric is effective for promoting inexperienced raters' scoring consistency since new teachers without rating experiences are still required to evaluate their students' open-ended tasks like writing essays. The inexperienced raters with a rubric were compared with those without. The experienced raters' scores of the sophomores that were required to take the writing classes for two consecutive semesters were compared qualitatively and quantitatively. The experienced raters' scores showed the students'improvement compared to when they took the first writing class, significant degradation before the second writing class began, and significant improvement at the end. Also, their pre-test essays in the second semester were better than those in the first semester and their post-test essays in the second semester were better than those in the first semester.

When the inexperienced raters evaluated the Korean English learners' essays with the same rubric, they could reflect the students' improvement as well as their initial writing proficiency as did the experienced raters. However, without the rubric, the inexperienced raters did not provide consistent scores compared to the students' writing proficiency. In other words, they scored the first two essays higher than the last two essays, which means that students might not receive legitimate scores for the better essays. Therefore, the rubric did help the inexperienced raters evaluate the English learners' essays consistently, reflecting their writing proficiency properly.

However, in terms of inter-rater reliability, the rubric helped very much for the inexperienced raters with the rubric. Their average scores for each test looked almost identical, which was not shown in the experienced raters. ANOVA results demonstrated that the inter-rater reliability of both the experienced raters and the inexperienced raters without the rubric was not secured. Their average scores for each test were very different from each other. It implies that a rubric might be effective to promote inter-rater reliability when raters evaluate the examinees' essays at the same time rather than when they evaluate them at different times even with the rubric. It is also needed to explore the effectiveness of

a rubric in terms of inter-rater reliability with the caution about when raters evaluate essays. Thus, workshops for raters should be held to promote their scoring consistency and to inform the effectiveness of a rubric to those who consider a rubric ineffective as described earlier. The samples of each score to activate their internal criteria based on a given rubric should be presented and teacher conferences are needed to evaluate students' essays together to promote scoring consistency within and among raters.

REFERENCES

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*(3), 371-383.

Beyreli, L., & Ari, G. (2009). The use of analytic rubric in the assessment of writing performance: Inter-rater concordance study. *Educational Sciences: Theory and Practice, 9*(1), 105-125.

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, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*(4), 587-603.

Cho, Seokhoon, Park, Juneon, Na, Kyunghee, Oh, Sehee, Kim, Changwon, Lee, Youngchan, & Eoh, Hyojin. (2008). *A study on introducing a new system of English speaking lecturers to schools*. Seoul: Ministry of Education.

Choi, Yeon Hee. (2002). FACETS analysis of effects of rater training on secondary school English teachers scoring of English writing. *Journal of Applied Linguistics Association of Korea, 18*(1), 257-292.

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.

Hyland, K., & Anan, E. (2006). Teachers' perceptions of error: The effects of first language and experience. *System, 34*(4), 509-519.

Jacobs, H. L., Zingkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House Publishers.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*(1), 81-112.

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

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

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

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

O'Loughlin, K. (1992). Do English and ESL teachers rate essays differently? *Melbourne Papers in Language Testing, 1*(2), 19-44.

Rezaei, A. R., & Lovorn, M. G. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39.

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

Silvestri, L., & Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. *International Electronic Journal of Health Education, 9*, 25-30.

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.

Weigle, S. C. (2002). *Assessing writing*. Cambridge, England: Cambridge University Press.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305-335.

Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching, 7*(1), 3-14.

APPENDIX

Writing Assessment Rubric

Assessment	Score	Response
Excellent	13, 14, 15	-Grammatically accurate; contains some minor errors; displays a broad range of grammatical structures; exhibits excellent use of compound and complex sentences -Completely coherent and well-organized; fully addresses the task; Writer's meaning is clear and the ideas are well developed. -Displays broad, sophisticated vocabulary
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 good use of compound and complex sentences -Generally coherent and well-organized; adequately addresses the task; Writer's meaning is generally clear and the ideas are fairly well developed. -Displays a relatively wide range of vocabulary
Average	7, 8, 9	-Fairly accurate; may contain occasional major errors; is somewhat limited to simple sentences; displays fair use of compound and complex sentences -At times incoherent and may contain parts that display unclear organization; fairly-adequate addressing of the task; Writer's meaning is at times obscure and the ideas mentioned could be more developed. -Displays somewhat narrow vocabulary
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 ineffective use of compound and complex sentences -Often incoherent and rather loosely organized; marginally addresses the task; Writer's meaning is often unclear and the ideas mentioned are rather inadequately developed. -Displays a simple and limited vocabulary
Poor	1, 2, 3	-Almost always grammatically inaccurate; displays only simple sentences -Generally incoherent; displays illogical, unclear organization; barely addresses the task; Writer's meaning is generally unclear. -Extremely limited and simple vocabulary

Examples in: English
Applicable Languages: English
Applicable Levels: Tertiary

Kyoung Rang Lee
English Language and Literature
Sejong University
209 Neungdong-ro, Gwangjingu
Seoul, 143-747
Tel: (02) 3408-3118
Email: kranglee@sejong.ac.kr

Received 23 March 2016
Revised 19 April 2016
Accepted 21 May 2016