Modern English Education, Vol. 15, No. 2, Summer 2014

An Investigation into Pre-Service and In-Service Teachers' Judgment of English Writing Performance*

Myo-Young Park**
Chonbuk National University
Jae-Woo Shim
Chonbuk National University

Park, Myo-Young & Shim, Jae-Woo. (2014). An investigation into pre-service and in-service teachers' judgment of English writing performance. *Modern English Education, 15*(2), 71-90.

The present study aimed to investigate rater characteristics on evaluating writing performance. Seven raters (three in-service teachers and four pre-service teachers) assigned scores 31 writings of Korean high school students. The writings were graded using a five-category analytic rating scale (Content, Organization, Vocabulary, Grammar, and Mechanics). The FACETS program revealed rater severity, reliability, and bias patterns among raters. Pre-service teachers, in general, scored leniently, while the most severe and lenient raters were all pre-service teachers. With respect to reliability, the result revealed that only pre-service teachers had rated the essays inconsistently. Also, they did not have high degrees of inter-rater reliability. In terms of bias, two pre-service teachers had a bias toward Mechanics and Vocabulary, while one in-service teacher toward only Mechanics. These results provide support for the notion that pre-service teachers need more rating exercises before they become in-service teachers.

[Writing assessment/Rater characteristics/Teaching experience/FACETS analysis/ 쓰기평가/ 평가자 특성/교수 경험/FACETS분석]

I. INTRODUCTION

Most Korean secondary school tests are high-stake tests, whether they are achievement tests or proficiency tests in that the test results may have some effects on

* This research was supported by the research funds of Chonbuk National University in 2013.
** First author: Myo-Young Park, Corresponding author: Jae-Woo Shim

their high school or college acceptance decision-making. Along with the increased portion of English performance assessments as part English achievement tests, valid and reliable measurements of students' performances have become an important issue. In secondary schools, students' performances of speaking and writing in English are assessed by their English teachers. However, depending on the teacher-related variables such as years of teaching experiences and experiences of rating performances, students' performances may be given unfair, unreliable test results. Thus, to help in-service teachers become better raters of writing performances, in particular, it is necessary that teachers are given rater training sessions that focus on scoring writing samples consistently in accordance with a set of writing traits. However, in fact, there have been very few rater training courses for English teachers in Korea. To name a few, an on-line rater training course has been offered by Education Broadcasting System (EBS) and an off-line course by each local Office of Education has been open. Since those rater training courses are not mandatory but optional at best, only a small number of English teachers have registered and taken the courses. To make things worse, after the Korean government decided not to pursue the administration of writing performance assessment known as National English Ability Test (NEAT), registrations in existing rater training courses have been reduced sharply. In the same vein, few rater training courses for pre-service teachers of English have been established in colleges as evidenced in course syllabus descriptions. Consequently, they end up facing the reality of having to assess writing performances of their students without adequate training.

Given the current practices of preparing English teachers for writing performances, this study aims to understand the scoring differences on written essays between in-service teachers and pre-service teachers and rater characteristics in terms of severity, reliability, and interaction effects.

II. LITERATURE REVIEW

1. Rater Experiences and Rater Training

Rater factors can be largely divided into raters' experiences of rating and raters' L1. Each factor has sub-components. For example, rating experiences comprise of teaching experiences writing skills, rating experiences, and rater training experiences. Almost all studies related to rater experiences have included rater training as a variable. In particular, studies that compared between trained (or experienced) and untrained (or novice) raters are easily found (Connor-Linton, 1995; Elder, Barkhuizen, Knoch, & von Randow, 2007; Knoch, Read, & Randow, 2007; Lumley 2005; Shaw & Weir, 2007;

Weigle, 1998; 2002). Weigle (1998) explored differences in rater severity and consistency between inexprerienced and experienced raters both before and after a rater training. 8 inexperienced and 8 experienced raters participated in scoring 60 writing essays. She found that inexperienced raters had a tendency to evaluate more severely and less consistently than experienced raters before training.

The research suggested that untrained raters were largely unreliable and were in need of rater training for improving their rater reliability (Elder et al., 2007; Lumley 2002, 2005; Shaw & Weir, 2007). However, a number of studies on rater training have showed low levels of reliability among trained raters themselves (Connor-Linton, 1995; Knoch et al., 2007; Lumley 2002, 2005; Weigle, 2002). Carrell (1995) mentioned that "no statistically significant effects for raters' training" (p. 175) and Kondo-Brown (2002) found that “Judgment of trained teachers raters can be self-consistent and overlapping in some ways, but at the same time, they may be idiosyncratic in other ways" (p. 25).

Very recently, Barkaoui (2011) and Lim (2011) also studied the differences in rater characteristics between inexperienced and experienced raters on writing assessment. Barkaoui found that the experienced raters assigned lower scores and gave more importance to linguistic accuracy than did the novice raters. Novice raters gave more importance to the efficiency of argumentation, exhibiting more variability. Lim conducted a longitudinal study and found that rating experiences may in some way related with rater severity and consistency. The above-mentioned studies overall supported the necessity for rater training.

Sweelder-Brown (1985) studied the rating differences in assessing essays, according to the amount of teaching experiences of faculty members. 6 trainer raters and 20 regular readers participated in the study. Of the 20 readers, 12 were instructors and eight were highly experienced graduate tutors, 5 of whom had been conducting their own writing development classes. All raters used two measuring instruments: a 1-6 holistic scale and a 1-4 analytic scale with 8 writing criteria. The result showed that highly experienced raters and trained raters had high levels of consistency between their holistic and analytic evaluations. This suggested that the rating reliability could be affected by the amount of experience and training. Another finding in the same study was that the more experience and training a rater had, the lower were both the holistic and analytic criteria scores. On the contrary, the inexperienced raters assessed the essays less critically than the experienced group. Concerning this phenomenon, Sweelder-Brown explained as follows.

> Perhaps they are afraid that they will make an error in judgment that will punish the student and are influenced by the notion that a lower grade is a value judgment of the student himself. Although we can only speculate, grades in this study with less experience and training were less critical of an essay's

qualities. (p. 55)

Despite the fact that some research studies have investigated raters' rating experiences as an important variable, there has been few studies dealing with any comparisons between in-service teachers and pre-service teachers in assessing English essays.

2. Many-facet Rasch Measurement (MFRM)

Many-facet Rasch Measurement is a family of Rasch models, which allow exploration of more than three facets of a performance test. MFRM can be conducted by FACETS, which is generally believed to one of the most appropriate computer programs for analyzing rater reliability. It provides estimates of examinee ability, rater harshness and consistence, rating category and scale difficulty, bias interactions between facets on a common log-linear metric or logit scale.

Eckes (2011) suggests that "the MFRM approach provides a rich set of highly efficient tools to account, and compensate, for measurement error, in particular rater-dependent measurement error" (p. 5). Accordingly, a number of studies have investigated various variables such as differences in rater severity and bias on oral performance tests (Bachman, Lynch & Mason, 1995; Bonk & Ockey, 2003; Brown, 1995; Hill, 1996; Iwashita, McNamara & Elder, 2001; Y. H. Kim, 2009; Lynch & McNamara, 1998; McNamara & Lumley, 1997; Weigle, 1998; Wigglesworth, 1997) and writing performance test (Eckes, 2008; Johnson & Lim, 2009; Knoch, Read, & Randow, 2007; Kondo-Brown, 2002; M. Y. Park, 2012; Shi, 2001; Y. S. Shin, 2010).

MFRM proposes a straightforward mathematical relationship among various facets that can affect on the result of the performance. In the present context where the facets are examinee, rater, and rating category, the relationship can be expressed as follow:

$$P = B - J - D - K$$

where

P = probability of a given score on a rating scale
B = ability of the student
J = severity of the judge
D = difficulty of the rating category
K = difficulty of a particular level on the rating scale

In this study, the following research questions were addressed;

1) What are the scoring differences (measure) on written essays between in-

service teachers and pre-service teachers?

2) What are the rater characteristics (severity, reliability, and interaction effects) of in-service teachers and pre-service teachers?

III. METHODOLOGY

1. Participants

1) Learners

The learners who participated in the present study were first year 31 high school students in Jeon-ju, Korea. Their first language (L1) was Korean. They have been learning English for 10 years. Their English proficiency ranged from intermediate to advanced level - their levels were based on the results of the mock tests by KICE (Korea Institute for Curriculum and Evaluation), although no official English scores were reported. All learners were asked to write for the same topic for 40 minutes.

2) Raters

Three in-service teachers and four pre-service teachers participated in assessing learners' written essays. The in-service teachers were all females and have more than five years teaching experience in Korean secondary schools. The pre-service teachers were English education majors in a University. 20 students that have taken a subject of "Assessment in English Education" for the sixth semester in 2013 fall. The four pre-service teachers out of a class of 20 pre-service teachers were solicited, who were taking as course titled as “Testing and assessment in English language”.

2. Rating Rubric

An analytic rating scale was used for assessing students' writing for the present study. The analytic rating scale has several aspects of writing or criteria rather than given as single score (Weigle, 2002). Some researchers suggested that analytic scoring is useful to train inexperienced raters (Weir, 1990) because separate scores can provide them with clear feedback including the strengths and weakness of their response (Linn & Gronlund, 2000). In addition, analytic scoring can be more reliable than holistic scoring (Huot, 1993).

The scoring rubric developed by M. Y. Park (2012) was used for the present study

(see Appendix A). The five categories of Content, Organization, Grammar, Vocabulary, and Mechanics were equally weighted, each category having a five-point Likert type scale. The reason was that this study was intended to understand the rater's assessing characteristics, not learner's ability.

3. Data Collection

Writing performance data consisted of essays produced by 31 Korean high school freshman. The students were given the same topic in order to avoid possible topic-type effect. The writing topic was "The domestic travel", which was a relatively familiar topic to them and was easy for them to describe their travel experiences without any content specific knowledge. The students asked to choose a place in Korea that they would like to recommend to foreigners.

The students had already been provided with information concerning the present study one week earlier. For example, the scoring rubric for the writing performance was given to them in advance. The writing test was conducted for 40 minutes. No dictionary was allowed. Any further information about the students were solicited because the present study was focused on raters, not the student examinees. Each of the 31 essays was scored by three in-service teachers and four pre-service teachers. The ratings by the raters were collected via mails or in person.

4. Data Analysis

The FACETS program (version 3.68.1; Linacre, 2011) was used to analyze the ratings by the raters. Three facets were applied to the program; student, rater, rating category. The present study focused on rater and rater-category measurement report.

After finishing the FACETS analysis, face to face interview sessions were held with the raters who showed some bias interactions with the rating categories. Since this interview in the present study was not used for in-depth qualitative approach, the responses by the raters were described in the form of summarization.

IV. RESULTS

1. FACETS Summary

The FACETS program indicates the examinees, raters, and rating category and the rating scale in the same logit scale. Figure 1 demonstrates the vertical ruler representing

these logit scales, which provides comparisons between facets, and within facets under consideration.

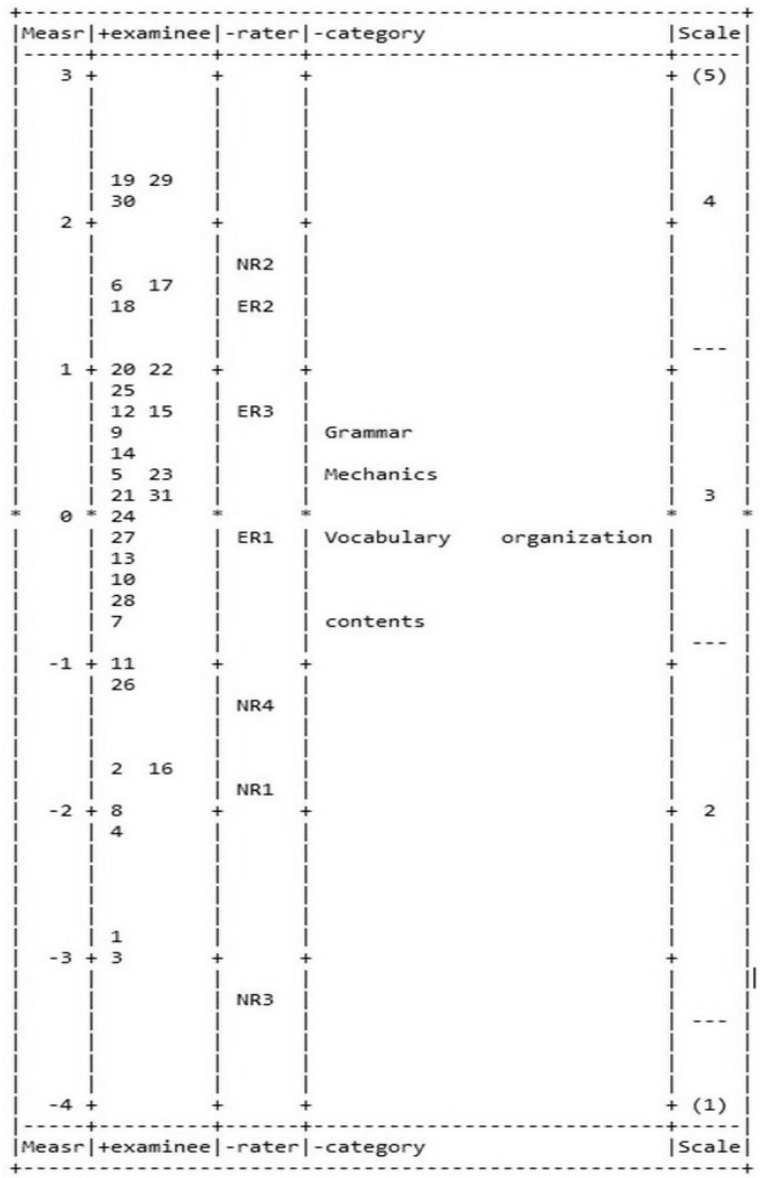


FIGURE 1 Variable Map from FACTES

Column 1 demonstrates the logit measures. Column 2 shows the ability of examinees' writings: on a continuum the top of the scale indicates higher ability, and the bottom of

the scale lower ability. For example, examinees 19 and 29 have the highest writing ability, while examinee 3 has the lowest ability. Value of logit 0 means average ability of examinees.

Column 3 shows rater severity. Raters with the value of plus logit are the ones who have evaluated the writing samples relatively severely; while raters with the value of minus logit are relatively lenient raters. Raters are evenly spread out from around + 2 logit to -2 logit. The most sever rater is NR2 (pre-service teacher 2) and the most lenient rater is NR3 (pre-service teacher 3). Interestingly, the most sever and the most lenient raters are pre-service teachers (NR1, NR2, NR3and NR4). Although ER2 is slightly severe, all in-service teachers have assessed neither severely nor leniently.

Column 4 demonstrates rating category difficulty; the higher the difficulty measure of a particular rating category, the more difficult for examinees to receive a high score on that category. Clearly, raters have evaluated the rating category of Grammar most severely and Contents most leniently.

Finally, Column 5 depicts the five-point rating scale that raters used to score examinees' essays. The three dotted horizontal lines across the column indicate the point at which the likelihood of getting the next higher rating begins to exceed the likelihood of getting the next lower rating for a given task. In other words, they indicate the scores of the writers at any given ability level on the scale are likely to receive (Weigle, 1998). For example, examinees between -4 and -3.5 logits are more likely to receive a score of 1. Those between about -3.5 and -1 logits are more likely to receive a score of 2.

2. Rater Characteristics

Table 1 provides information about raters who participated in the present study; Column 1 raters, Column 2 observed overage and Column 3 fair-M average. Observed average (mean: 3.2) is calculated by dividing the observed score by the observed count. Fair-M average is considered more accurate mean value (mean: 3.15), which is calculated from the values of other facets. Thus, examinees are given slightly higher scores than the scores FACETS program estimates. This Fair-M average should be used in high-stakes tests including college entrance exams.

TABLE 1

Rater Measurement

Rater	Obsvd Average	Fair-M Average	Model. Measure	S.E.	Infit MnSq	Infit ZStd	Exact agree Obs%	Exact agree Exp%
NR3	4.4	4.53	-3.25	.15	1.37	2.5	18.1	17.5
NR1	3.8	3.88	-1.85	.12	1.67	5.1	25.0	25.3
NR4	3.6	3.61	-1.36	.12	.54	-5.1	24.1	27.5
ER1	3.1	3.02	- .21	.12	.98	- .1	31.2	29.8
ER3	2.6	2.59	.71	.12	.56	-4.7	28.7	28.1
ER2	2.3	2.28	1.39	.12	1.12	1.0	25.6	25.0
NR2	2.2	2.16	1.67	.13	.72	-2.6	23.8	23.3
Mean	3.2	3.15	- .41	.12	.99	-.6		
S.D.	.8	.89	1.82	.01	.43	3.8		

RMSE .12 S.D. 1.82

Separation 14.61 *Reliability* 1.00

Fixed (all same) chi-square 1161.4 *d.f.* 6 *significant (probability)* .00

Severity Column 4 in the Table 1 shows the rater severity measure. The measure shows accurate value of severity. The most lenient rater is NR3 (- 3.25) while the most lenient rater is NR 2. NR3 (-3.25), NR1 (-1.85), NR4 (-1.36), and ER1 (-.21) have assigned a score relatively severely. ER3 (.71), ER2 (1.39), and NR2 (1.67) have evaluated the writing samples relatively leniently. This rater measure report shows that NRs (pre-service teachers) have made extreme decisions.

Intra-rater reliability Next Column in the Table 1 is the infit measure and Z-score, which explains intra-rater reliability. This study follows fit indices of McNamara (1996), which ranges from .75 to 1.3. Fit indices show the degree to which observed ratings match the expected ratings that are generated by the Rasch model. The fit value greater than 1.3 shows more variation than expected in the ratings called *misfit*, while the fit value below the .75 indicates the other variation called *overfit*. When infit contradicts Z-score(±2), it is natural to follow the Z-score. NR3 (2.5) and NR1 (5.1) are misfit while NR4 (-5.1) and ER3 (-4.7) are overfits. Overall, pre-service teachers have more variation than in-service teachers. TABLE 2 gave a typical example to show why NR1 and NR 4were misfit and overfit respectively. Sample 27, 36 and 43 were selected because the two raters demonstrated a striking rating difference.

TABLE 2
Examples of Ratings

Category	Sample 27		Sample 36		Sample 43	
	NR1	NR4	NR1	NR4	NR1	NR4
Content	5	3	1	3	4	4
Organization	5	3	1	3	3	5
Vocabulary	5	3	4	3	3	4
Grammar	4	3	4	3	5	4
Mechanics	1	3	2	3	3	4

The most misfit rater (NR1) and the most overfit rater (NR4) were interviewed to determine the cause of their abnormal scoring. NR1, a misfit rater, answered that she had took more than three days to evaluate test takers' essays. According to Eckes (2011), rater misfit can indicate an idiosyncratic rating style or otherwise overly inconsistent rating behavior (p. 60). Overall, NR1 was not consistent with her assessment of the essays (see Table 2). NR4, an overfit rater, mentioned that he would use a rating scale 3. His remark suggests the central tendency toward the mean or the halo effect. A rater exhibits a central tendency

> when he or she overuse the middle category, or middle categories, of a rating scale while assigning fewer scores at both the high and low ends of the scale. ... result in ratings that overestimate the proficiency for low-performing examinees and underestimate the proficiency for high-performing examinees. (Eckes, 2011, p.62)

NR4 used one or two scale points in 28 out of 31 writing samples. In addition, NR4 used only one scale point to 5 writing samples; for example, sample 27 (3, 3, 3, 3, 3) and 36 (3, 3, 3, 3, 3). In addition, he confessed that he becomes easily tired of doing things. Based on the interview, it can be said that his personality was reflected on the scoring. Thus, it could be said that his personality would be reflect on assigning scores (see the scoring examples of writing samples in Appendix B)

Inter-rater reliability Table 1 also shows the information of inter-rater reliability, providing observed exact agree % and expected exact agree %. Exact agreement observed % indicates a range of agreement among raters and exact agreement expected % shows a range of agreement to Rasch model. If Obs% is approximately equal to Exp%, then the raters would behave like independent experts; if Obs% is much

larger than Exp%, then the raters would score accurately like scoring machines (Linacre, 2011). The difference (2.6) between Obs% and Exp% of in-service teachers is small, while the difference (8.1) between the two % of pre-service teachers is significant large (see Table 3).

TABLE 3

Comparison of Inter-rater Reliability between Two Groups

	Obs%	Exp%	Inter-rater reliability
Pre-service teachers (NRs)	91	99.1	-26.1
In-service teachers (ERs)	85.5	82.9	.15

Linacre (2011) suggested calculating inter-rater reliability using the following Rasch version of the kappa index:

$$\text{Rasch-kappa} = (\text{Obs\%} - \text{Exp\%}) / (100 - \text{Exp\%})$$

According to the Rasch model, if the Rasch-kappa value much greater than 0, it means overly high inter-rater agreement, while if the value has a negative Rasch-kappa value, it means much less inter-rater agreement. Thus, it can be assumed that pre-service teachers in the present study do not have much inter-rater agreement (-26.1) and on the contrary in-service teachers have a significantly high agreement among them (0.15).

3. The Difficulty and Consistency for the Rating Categories

Table 4 presents the difficulty measurement report for the five rating category facets. According to the measure of difficulty(logit), Mechanics and Grammar are relatively difficult, compared to Content, Vocabulary, and Organization. The most leniently scored category is Content (-.69) and the most severely scored category is Grammar. Although the difficulty spans between these two categories of Content and Grammar were small as 1.25 logits, reliability separation index is quite high (.94). Also the chi-square of 85.5 with 4 d.f. was significant at $p<.00$. Thus the null hypothesis that all categories would be equally difficult was rejected.

This Table also provides information concerning how consistently the categories have been assessed. Infit values put a range from .75 to 1.3. Although Vocabulary, Mechanics, and Grammar are slightly overfitted, the span is not significant. Accordingly, the overall patterns between categories are consistent.

TABLE 4
Rating Category Measurement Report

N	Category	Difficulty(logit)	Error	Infit(MnSq)
1	Content	-.69	.11	1.04
3	Vocabulary	-.13	.10	.86
2	Organization	-.09	.10	1.17
5	Mechanics	.35	.10	.85
4	Grammar	.56	.10	.94
	Mean	.00	.10	.97
	SD	.48	.00	.13

Separation reliability .94

Fixed (all same) chi-square 85.5, *d.f.* 4, *significant* $p<.00$

4. Rater – Rating Category Bias Interaction

Bias analysis is to investigate any interaction among particular facets. Raters would have a particular pattern in scoring examinees' performance among other interactions such as rater-examinee interactions, rater-scale interactions, or rater-task type interaction. The present study has focused on bias interaction between raters and rating categories.

TABLE 5
Bias Interactions between Raters and Rating Categories

Rater	Category	Obs-Exp average	Bias(logit)	Error	z-Score
NR1	Mechanics	-.26	-.55	.26	-2.12
ER1	Mechanics	.30	.63	.26	2.45
NR3	Vocabulary	.30	1.34	.45	2.95

Table 5 shows the information of rater-category bias. Generally z-score greater than +2 or smaller than -2 are considered significant bias (McNamara, 1996). Two pre-service teachers (NR1 and NR3) and an in-service teacher (ER1) had significant bias interaction with rating categories. NR1 and ER1 have exhibited bias interactions for Mechanics and NR3 for Vocabulary. Obs-Exp average is an index of bias. NR1 who have below -2 *z*-score (-2.12) is a systematically more lenient behavior than is normal for the rater in question. In contrast, ER1(*z*-score 2.45) and NR1(*z*-score 2.95) is more severe behavior for Mechanics and Vocabulary relatively.

V. DISCUSSION AND CONCLUSION

The results of the present study suggested that inexperienced raters may have be less consistent than experienced raters in scoring the writing samples. In-service teachers and pre-service teachers of Korean secondary schools participated in grading high school students' writing samples with a 5-category analytic rubric. It was found that pre-service teachers were biased on both side-weighted. In other words, their severity measure showed some variability as indicated by -3.25, -1.85, -1.36 and 1.67, while among in-service teachers, severity measure came together to be zero. It was also found that he pre-service teachers scored the writing samples relatively leniently. This result was quite opposite to the study by Weigle (1998), who found that inexperienced raters tended to be more severe.

In terms of intra-rater consistency, all pre-service teachers and one in-service teacher showed inconsistency on scoring the writing samples. Specially, one of the pre-service teachers revealed a serious central tendency "by avoiding scores at the end of the scale" (McNamara, 1996, p.124). According to the interview with the rater, however, the reason was different from what McNamara claimed. The rater had a tendency of getting tired too easily and thus overlooked most of his rating of the essays. This finding suggests that a rating trainer should identify raters who have a central tendency and discuss the importance of fair and reliable performance assessment.

With regard to inter-rater consistency, pre-service teachers also showed serious disagreement on grading the students' products among the group members. This suggests that pre-service teachers should be given more opportunities to rate essays. They may be introduced to concepts of measurement, construct of writing performance, issue of validity and reliability, different types of rubrics, and so forth.

Concerning bias interactions between scoring categories, all raters generally interpreted the notions of the scoring categories and applied them to the task of assessing the writing samples appropriately. However, more detailed criteria for assessing Mechanics component may be needed for both pre-service and in-service teachers of English because some of them seemed to have been inconsistent with Mechanics. M. Y. Park (2012) found a similar result. In her study, Korean in-service teachers of English had some difficulty in understanding and interpreting the categories of Mechanics and Vocabulary. Overall, the findings of this study are similar to those studies that investigated the differences in rating patterns between novice and experienced raters (Kobayashi & Rinnert, 2003; Wolfe, Kao, & Ranney, 1998).

This study has some limitations as well. Most of all, this research findings of this study are confined to general aspect of writing assessment such as severity, consistency, bias interactions between two rater groups of pre-service teachers and in-service teachers.

To understand more specifics of rater characteristics, it would be necessary to include more assessment variables and involve more raters as subjects. In addition, the findings of this study may be corroborated or unsupported through other qualitative analysis techniques. For examples, a think-aloud protocol may play a crucial role in interpreting raters' decision-making processes. Another limitation of this study is concerned with time limits given to raters. The variability in completing scoring the essays differed from raters to raters. More constraints on the time limits would have eliminated the phenomenon of inconsistency due to the unlimited time span that lasted up to one week.

REFERENCES

Bachman, L. F., Lynch. B., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-256.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51-75.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

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

Carrell, P. L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing, 2*(2), 153-190.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*(4), 762-765.

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

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement.* Frankfurt am Main: Peter Lang.

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

Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing, 5*(2), 29-49.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating students essays. In M. A. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206-236). Cresskill, NJ: Hampton.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(2), 401-436.

Jonson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485-505.

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

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

Kobayashi, H., & Rinnert, C. (2003). Coping with high imposition requests: High vs. low proficiency EFL students in Japan. In Flor, A. M., Juan, E. U., & Guerra, A. F. (Eds.), *Pragmatic competence and foreign language teaching* (pp. 161-184). Universitat Jaume I: Spain.

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

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560.

Linacre, J. M. (2011). *A user's guide to WINSTEPS: Rasch-model computer programs.* Chicago: Winsteps.com.

Linn, R., & Gronlund, N. (2000). *Measurement and assessment in teaching*. Prentice Hall, New Jersey.

Lumley, T. (2002*).* Assessment criteria in a large scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276.

Lumley, T. (2005*). Assessing second language writing: The rater's perspective.* Frankfurt am Main: Peter Lang.

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

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

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140-156.

Park, Myo-Young. (2012). Exploring the raters' bias on a EFL writing assessment using Multi-faceted Rasch Measurement. *Studies in English Education, 17*(2), 178-202.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in*

assessing second language writing. Cambridge, UK: Cambridge University Press.

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.

Sweelder-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluations. *English Journal, 75*(5), 49-55.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-267.

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

Weir, C. J. (1990). *Communicative Language Testing*. Hemel Hempstead: Prentice Hall.

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

Wolfe, E., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4), 465-492.

APPENDIX A

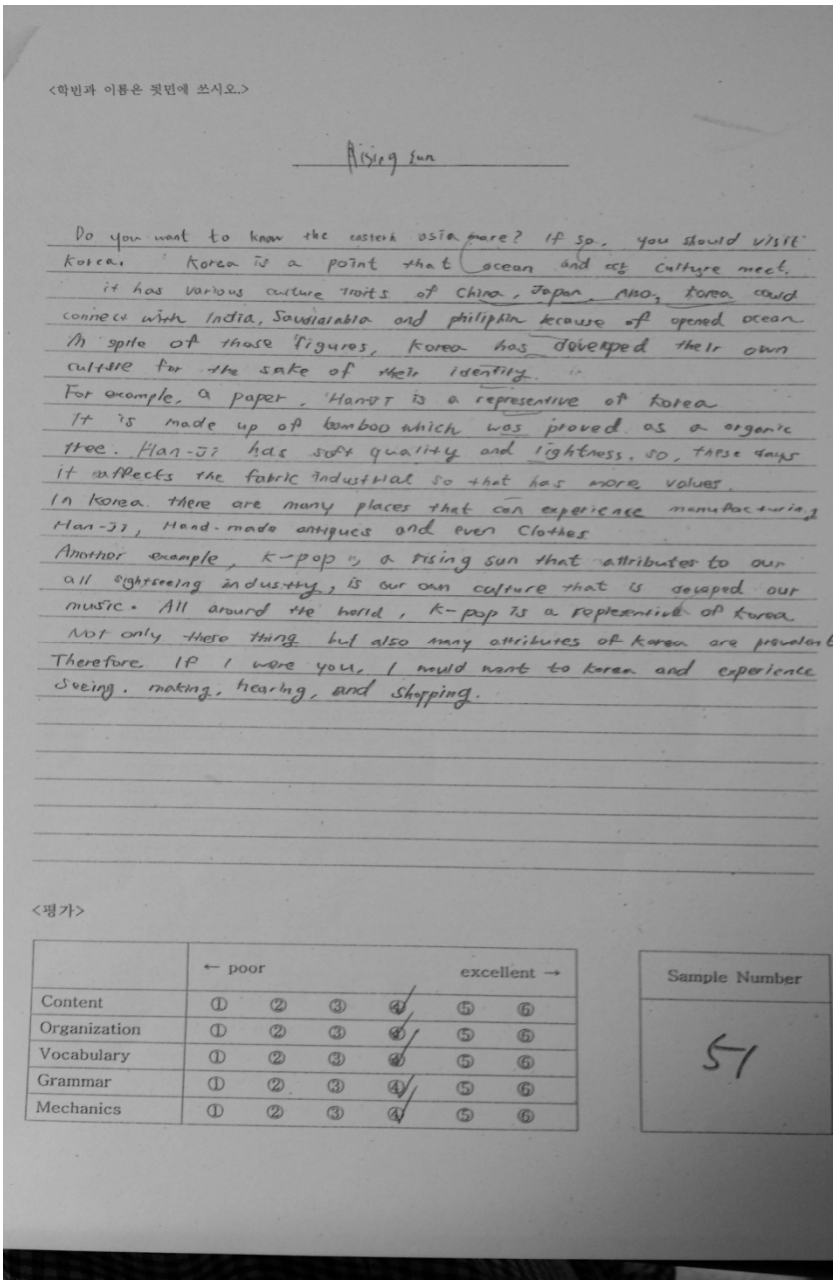
The Rating Scale

5: Advanced **4**: High **3**: Intermediate **2**: Low **1**: Novice

Assessment criteria		Rating scales	Criteria
Content	one clear topic; logical development; main idea; supporting details	5	There is one clear topic; Effectively addressed topic and task, using clearly appropriate explanations; sufficient supporting details.
		4	There is one clear topic; Well addressed topic and task using appropriate explanations, sufficient supporting details
		3	There is one clear topic; Addressed topic and task using developed explanations, somewhat insufficient supporting details
		2	There is not one clear topic. Limited development in response to the topic and task using inappropriate explanation, lack of supporting details
		1	There is no single topic. very limited or no development in response to the topic and task using no explanation, no supporting details.
Organization	The sequence of introduction, body, and conclusion; the cohesion	5	Well organized and cohesive devices effectively used
		4	Fairly well organized and cohesive devices adequately used
		3	Loosely organized and in complete sequencing; absent cohesive device used
		2	Disconnected and lack of logical sequencing inadequate order of ideas
		1	No organization and no use of cohesive devices
Vocabulary	word choice	5	Appropriate, accurate, and natural choice of words
		4	Relatively appropriate, accurate, and natural choice of words
		3	Adequate choice of words but somewhat unnatural sometimes
		2	Limited and confused use of vocabulary
		1	Very limited vocabulary
Grammar	sentence-level structure; use of relative clauses. modals, articles, verb forms, agreements and tense sequencing	5	Almost no errors, full control of syntactic variety
		4	Almost no errors, good control of syntactic variety
		3	Some errors, fair control of syntactic variety
		2	Many errors, poor control of syntactic variety
		1	Severe and persistent errors, no control of syntactic variety
Mechanics	Punctuation; spelling; capital letter	5	Perfect spelling and punctuation and good use of capitals
		4	Few errors in spelling, punctuation and capitals
		3	Some errors of spelling, punctuation and capitals
		2	Serious problems with spelling, punctuation and capitals
		1	severe problems with spelling, punctuation and capitals

APPENDIX B

Sample 51 Rated by NR4



<학번과 이름은 윗칸에 쓰시오.>

Rising Sun

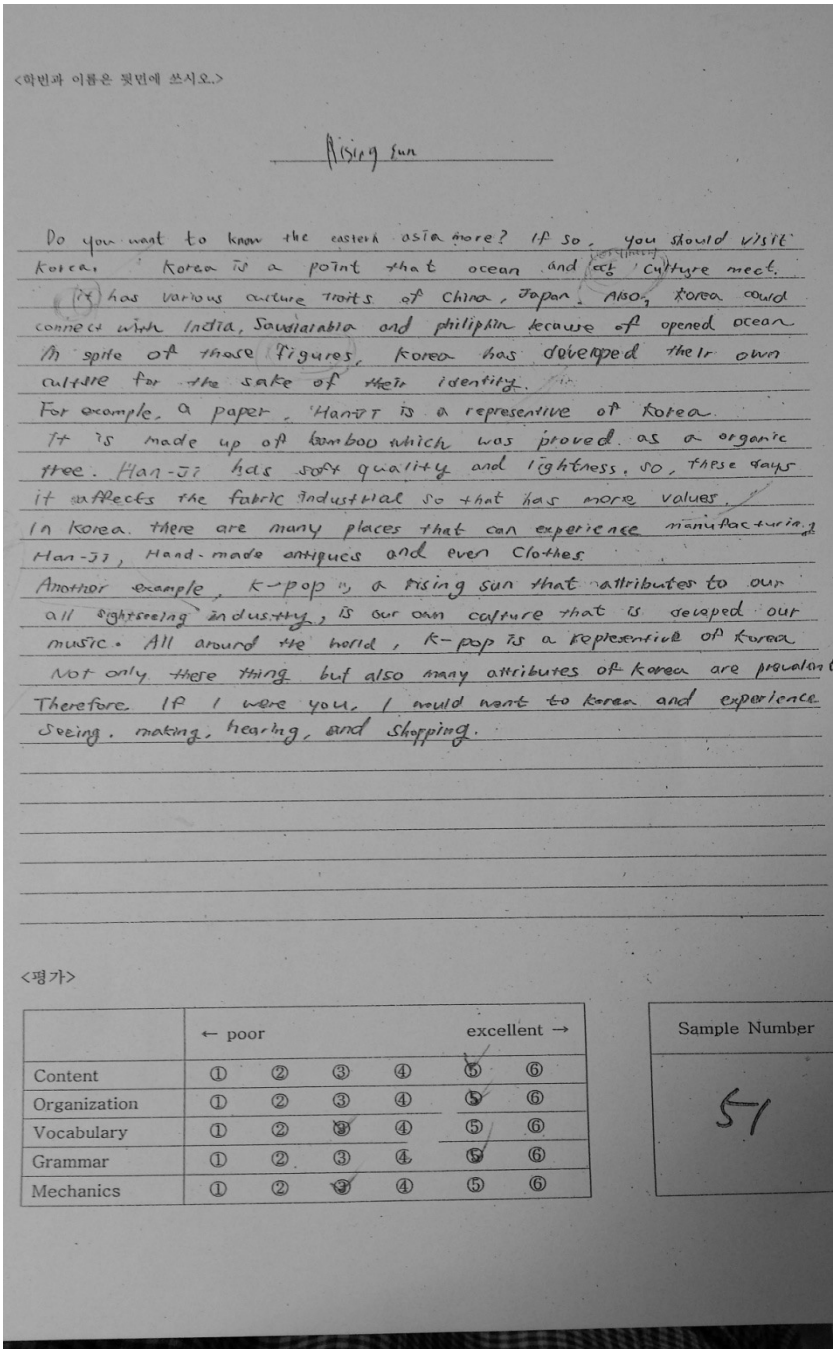
Do you want to know the eastern asia more? If so, you should visit Korea. Korea is a point that ocean and culture meet.
it has various culture traits of China, Japan, Also, korea could connect with India, Soudiarabia and philipin because of opened ocean. In spite of those figures, Korea has developed their own culture for the sake of their identity.
For example, a paper, 'Hanji' is a representive of korea. It is made up of bamboo which was proved as a organic tree. Han-ji has soft quality and lightness. so, these days it affects the fabric industrial so that has more values.
In korea there are many places that can expericase manufacturing Han-ji, Hand-made antiques and even Clothes.
Another example, K-pop is a rising sun that attributes to our all sightseeing industry, is our own culture that is developed our music. All around the world, K-pop is a representive of korea. Not only these thing but also many attributes of korea are prevalent. Therefore. If I were you, I would want to korea and experience Seeing, making, hearing, and shopping.

<평가>

	← poor					excellent →
Content	①	②	③	④	⑤	⑥
Organization	①	②	③	④	⑤	⑥
Vocabulary	①	②	③	④	⑤	⑥
Grammar	①	②	③	④	⑤	⑥
Mechanics	①	②	③	④	⑤	⑥

Sample Number
51

Sample 51 Rated by NR1



<아빈과 이름은 뒷면에 쓰시오.>

Rising Sun

Do you want to know the eastern asia more? If so, you should visit Korea. Korea is a point that ocean and ect culture meet. It has various culture traits of China, Japan. Also, Korea could connect with India, Saudiarabia and philipin because of opened ocean In spite of those figures, Korea has developed their own culture for the sake of their identity.

For example, a paper, Hanji is a representive of Korea. It is made up of bamboo which was proved as a organic tree. Han-Ji has soft quality and lightness, so, these days it affects the fabric industrial so that has more values. In Korea there are many places that can experience manufacturing Han-Ji, Hand-made antiques and even Clothes.

Another example, K-pop is a rising sun that attributes to our all sightseeing industry, is our own culture that is delaped our music. All around the world, K-pop is a representive of Korea Not only these thing but also many attributes of Korea are popular Therefore If I were you, I would want to Korea and experience Seeing, making, hearing, and shopping.

<평가>

	← poor				excellent →	
Content	①	②	③	④	⑤	⑥
Organization	①	②	③	④	⑤	⑥
Vocabulary	①	②	③	④	⑤	⑥
Grammar	①	②	③	④	⑤	⑥
Mechanics	①	②	③	④	⑤	⑥

Sample Number
51

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

Myo-Young, Park
Dept. of English Education, Chonbuk National University
567 Baekje-daero, deokjin-gu, Jeonju-si, Jeollabukdo 561-756, South Korea
C.P.: 010-9152-5239
E-mail: usgirl2002@naver.com

Jae-Woo, Shim
Dept. of English Education, Chonbuk National University
567 Baekje-daero, deokjin-gu, Jeonju-si, Jeollabukdo 561-756, South Korea
Tel. (063) 270-2729
E-mail: shimjw@jbnu.ac.kr

Received 11 March 2014
Revised 9 May 2014
Accepted 16 May 2014