Comparisons of In-service and Pre-service Korean Language Teachers' Writing Assessments Using an Empathizing Task

Jang, Eun Ju Sang-am middle school

Park, Young Min Korea National University of Education

I. Introduction

Teachers are stressed by various factors, and assessment is the main factor that has been emphasized of late. Assessment must accurately evaluate the level of knowledge or skill learned by students, and must capture and explicitly reveal the results of learning taking place in the classroom. Despite these, it is however more difficult to assess writing skill than it is to assess other domains of Korean language education. Writing ability is widely considered to be the primary feature of the writing domain, and should be evaluated through direct assessment. To measure writing ability, Korean language teachers should evaluate writing objectively, even when essays are written subjectively. To date, only minimal training in writing assessment has been offered to pre-service and in-service teachers.

In recent years, there have been various studies of writing assessment, and especially of raters. In keeping with these previous studies, this paper shall examine Korean language teachers as raters through a comparison of the writing assessments performed by in-service and pre-service teacher, and make suggestions how to further address additional needs in rater training. For this study, an empathizing writing task was selected. Although expressive writing is often used in

Korean language class, it is difficult to rate reliably in comparison to expository or argumentative writing. Therefore, this study aims to compare writing assessment of pre-service and in-service Korean language teachers using an expressive writing empathizing task.

II. Literature Review

1. Rating experience

Acting as a rater is a critical component of writing assessment carried out by Korean teachers. When a Korean teacher scores student work, writing assessments are often affected by the teacher's beliefs and attitudes about writing and existing, as well as their knowledge of writing assessment. The teacher's gender and teaching experience may also be a factor.

As compared with pre-service teachers, in-service teachers have likely acquired more knowledge about the design and practice of writing assessment. Mertler (2005) compared the assessment literacy of pre-service and in-service teachers using his Classroom Assessment Literacy Inventory (CALI). In his study, in-service teachers performed best on administering, scoring, and interpreting the results of assessments (Standard 3), and performed these tasks far better than did pre-service teachers. In-service teachers, however, performed worst on developing valid grading procedures (Standard 5), matching the performance of pre-service teachers much more closely.

Y. Park (2013) reported on the writing assessment knowledge of Korean teachers. In his study, in-service teachers had statistically significant higher scores than did pre-service teachers. He suggested that knowledge of writing assessment might be increased gradually through teacher training courses as well as by time spent in the teaching profession. It stands to reason that teachers' rating skills would improve with experience.

Based on teaching experience, differences emerge in how raters are affected by student writing and on what raters focus on during writing assessments. Generally, it seems that the more experienced raters have, the higher students are scored (Keech & McNelly, 1982; Song & Caruso, 1996). In a comparison of assessment performance between in-service and pre-service Korean teachers, Park and Choi (2009) reported a high correlation between the grades given by both groups, and that pre-service teachers graded more severely based on scoring criteria than did in-service teachers. Additionally, in-service teachers focused more on content, while pre-service teachers focused on both content and style in writing assessment.

In-service teachers showed higher writing assessment efficacy than did pre-service teachers, and pre-service teachers with teaching practice experience showed higher assessment efficacy than pre-service teachers without teaching experience (Park, 2010, 2011). In an examination of the performance of in-service Korean teachers, Park (2011) reported that there was no significant difference in the way some teachers regarded their own performance based on gender or career length.

Given these results, it is difficult to say definitively that a teacher's degree of experience strengthen his or her skills as a rater. Harari and McDavid (1973) asked in-service teachers and sophomores to rate essays labeled with "desirable" and "undesirable" names as a means of discovering any gender stereotype assumed by the rater. Student writings labeled with "desirable" names were generally rated higher by teachers than writing labeled with "undesirable" names. Undergraduate students studying to be teachers, however, did not rate student work based on student names. In addition, they concluded "experienced teachers accumulate these stereotypical expectations and biases over time with their training and experience as teachers" (Harari & McDavid, 1973: 225). Therefore, teachers as raters should examine whether the criteria they use is suitable for the writing task, reflect on the construct and check the rating process.

2. The effects of gender and the other variables

There are two kinds of study related to gender in writing assessment: (1) writing assessment based on the rater's own gender, and (2) differences based on the rater's perception of the writer's gender.

Female Korean teachers tend to grade more severely than male Korean teachers, but the difference between the rater's gender and his or her perception of student gender is not significant (Park & Choi, 2009). Conversely, other studies have reported that female teachers grade writing samples more leniently than male teachers (Bernard, 1979; Jeung, 2011). In a sample writing task, male teachers graded general explanatory and argumentative essays, while female teachers graded book reports. The gender of rater may affect the writing score, as one gender may grade more severely or leniently depending on the genre or writing task being performed.

Peterson and Kennedy (2006: 42) revealed that "teachers' assessments of the quality of the writing were often influenced by their perceptions of the writer's gender." Whether the rater recognized the examinee's gender as being the same as his/her own could affect the assessment of student writing. Roen (1992) found that teachers evaluated same-sex-named essays higher than other-sex-named essays. Peterson (1998), however, suggested that teachers tend to write more comments on essays written by students who share their gender. Male teachers suggested the need for review more when assessing the writing of boys compared to girls, and tended to praise girls' writing more than that of boys, while female teachers did the reverse (Etaugh et al., 1988; Haswell & Haswell, 1996). If raters believe that the writers share their gender, they tend to assess the writing more severely or negatively. This is called *same-sex depreciation* (Haswell & Haswell, 1996).

In addition to, rater and examinee gender, writing styles, such as writing order, writing level, writing media and readability (Whithaus et al., 2008), and topic and genre (Bouwer, 2014), may also influence writing assessment. According to the genre of assigned writing task,

the pattern of writing assessment might vary. Wiseman (2012) reported that raters scored narrative essays more severely than argumentative essays. J. Park (2013) found that while some Korean language teachers were a good fit in consistency of narrative writing rating but overfit in consistency, lack of discrimination, of argumentative. Other teachers were doing the reverse. Therefore, Korean language teacher have to be well-informed of the rubric classified by genre.

III. Method

1. Data collection

Essays were collected from two high schools in Gyeong-gi Province, Korea. Students were asked to read an essay titled 『*Musoyu*』 (Non-possession) written by the late Buddhist monk, Beopjeong. He wrote an anecdote in which he gave an orchid to someone after realizing that he had grown too attached to it. In present study, the assigned writing task was the following;

> "*Imagine that you are the orchid in this essay, and then write a letter to the writer (Beopjeong) expressing your feelings.*"

For this writing task, students had to understand the relevance of the object (the orchid) and empathize with it. This task was designed for use in writing classes[1] with the intent of also serving as a means of promoting character education.

Fifty writing samples were scored after disqualifying incomplete

1 Although previous researches of empathy have used actors or fictional characters as the targets of empathy, it is necessary to choose a person spontaneously relating his or her own real experience when examining empathic capability (Klein & Hodges, 2001). Everyone is capable of growing attached to something or developing an obsession with an object.

writing samples or essays not submitted by the deadline. However 13 samples did not meet the task requirements. Therefore, 37 writing samples (19 by boys and 18 by girls) were examined statistically.

2. Participants

Participants were 83 teachers. A total of 35 were pre-service teachers with no teaching experience, and 48 were in-service teachers who were currently working all over the country. Twenty-seven were male and 56 were female.

Table 1. Participants

	Pre-service	In-service			Total
		Under 5 years	5-10 years	11-20 years	
Female	24	9	15	8	56
Male	11	5	5	6	27

In-service teachers' experience varied; 14 teachers had less than 5 years of experience, 20 teachers had 5 years to 10 years of experience, and 14 teachers had 11 years to 20 years of experience. In addition, 35 pre-service teachers were juniors at Korea National University of Education. These were enrolled in a writing instruction class but had never experienced writing instruction or assessment.

3. Procedure

Writing assessment packages including the writing prompt and grading rubric (table 2) were forwarded to the raters, along with the written essays from which identifying student information had been removed. Pre-service teachers graded their papers in May 2014, and in-service teachers graded theirs in August 2014. The rubric used by raters provided an "essay writing rubric" (Bae, 2010) and a "narrative

essay rubric" (Park & Park, 2011). This focused on *Content, Organization, Expression, Word choice,* and *Style and mechanics*. Raters were asked to use full 6-point scales throughout the scoring process.

Table 2. Rubric

category	Criteria
Content	The writer offered a clear topic that might attract the audience's attention. The writer offered details that support the topic.
Organization	The writer sequenced coherently and cohesively. The writer created a piece that is easy to follow.
Expression (tone and attitude)	The writer constructed sentences that make the essay interesting and original. The writer used obvious and understandable expressions with their own voice.
Word choice	The writer selected words with precision.
Style and mechanics	The writer formed grammatically correct style; mechanics, punctuation, spelling, and splitting a paragraph.

The analysis was carried out using FACETS ver 3.71.3. Seven facets were set up: (1) writing samples, (2) examinee gender, (3) raters (both in-service and pre-service), (4) rater's gender, (5) rater's career length, (6) whether the rater was in-service or pre-service, and (7) grading criteria. The rating system calculated the following: (1) severity and consistency of the rater's grading, (2) bias due to career length and grading criteria, and (3) any bias rooted in examinee gender and rater gender.

IV. Results and Discussion

1. Preliminary analysis

Figure 1 graphically depicts the manner in which rater severity and leniency are captured by many-facets Rasch model (MFRM) analyses. This figure shows the distribution of rater severity measures

and the distribution of examinee performance measures[2]: (1) the "examinee" column shows writing samples; (2) numbers 1~48 reflect in-service teachers and 49~83 reflect pre-service teachers in the "rater" column; (3) in the "career" column, "1" corresponds to pre-service teachers without teaching experience, "2" refers to teaching experience of less than 5 years, "3" indicates 5~10 years of teaching experience; and "4" denotes 11~20 years of teaching experience; and (4) position as a pre-service teacher or in-service teacher.

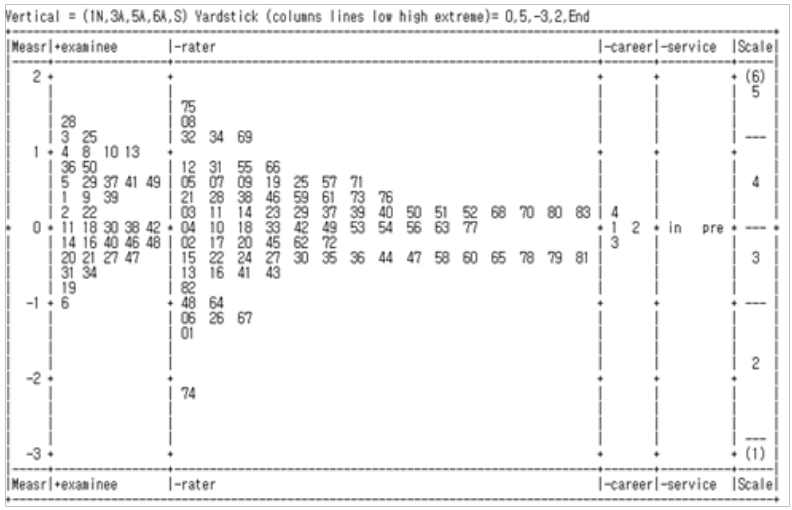

Figure 1. Examinee-Rater-Career-Service Map

1) Severity

Rater severity ranged from –2.22 logit to 1.64 logit. Both the most severe and lenient raters were pre-service teachers. Rater severity among in-service teachers ranged from –1.49 logit to 1.38 logit. The distribution of in-service teacher severity varied less than that among pre-service teachers. Experience with writing assessment might ex-

2 For the sake of convenient reference, three facets were excluded in the vertical ruler and their statistics are presented in the next chapter: student's gender, rater's gender, and criteria.

plain this.

As table 3 shows, the difference between in-service and pre-service teachers' severity was 0.06 logit. As indicated by the chi-square value, the difference of whether teachers were in-service or pre-service was statistically significant.

Table 3. Service Measurement Report (arranged by MN)

Service	Obsvd	Fair(M)		Model	Infit		Outfit		Correlation	
	Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	PtMea	PtExp
in	3.74	3.68	-0.03	0.01	1.07	4.9	1.08	5.6	0.65	0.65
pre	3.61	3.63	0.03	0.01	0.89	-6.5	0.90	-5.9	0.64	0.65
RMSE .01 Adj (True) S.D. 0.04				Separation 3.05			Strata 4.40		Reliability 0.90	
Fixed (all same) chi-square: 10.3				d.f.: 1			significance (probability): 0.00			

Compared to pre-service teachers, in-service teachers tended to score more leniently. This result supports previous research. When scoring writing samples, in-service teachers considered students' development levels, whereas pre-service teachers evaluated them according to adult writing levels and therefore rated more severely. This suggests that pre-service Korean teachers need to have the opportunity to understand students' writing development and rate writing samples.

Severity is usually attributed to not only writing rating experience, but also gender. In present study, raters were analyzed as 4 group of gender; In-service Female, In-service Male, Pre-service-Female, Pre-service Male.

Table 4 shows that female teachers were more lenient than male teachers. Female in-service teachers scored most leniently while male pre-service teachers scored most severely. In previous study, female teachers usually score leniently than male. However, the results were reverse in present study. Empathy is often examined in relation to gender; females are more empathic than males (Toussaint & Webb,

Table 4. Rater Gender Measurement Report (arranged by MN)

Rater Gender	Obsvd	Fair(M)		Model	Infit		Outfit		Correlation	
	Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	PtMea	PtExp
In-F	3.82	3.75	-0.09	0.01	1.10	5.9	1.12	6.6	0.66	0.65
Pre-F	3.66	3.67	-0.02	0.02	0.92	-3.8	0.93	-3.2	0.67	0.68
In-M	3.56	3.61	0.04	0.02	1.00	0.0	1.00	0.2	0.63	0.63
Pre-M	3.51	3.59	0.07	0.02	0.83	-5.9	0.83	-5.9	0.55	0.53
RMSE .02 Adj (True) S.D. .10				Separation 5.23			Strata 7.31		Reliability .96	
Fixed (all same) chi-square: 80.7				d.f.: 3			significance (probability): .00			

2005; Schulte-Rüther et al., 2008; Klein & Hodges, 2001) The result of this study suggests that writing assessments that incorporated the empathizing task were much more likely to be influenced by rater gender than by teaching experience.

2) Consistency

Generally, infit mean square values greater than 0.75 and less than 1.3 are considered fitting. Further, infit mean square values less than 0.75 are overfitting and greater than 1.3 are misfitting (McNamara, 1996). Among all 83 raters, fit rater consistency was 66.26% (55), while overfitting raters comprised 20.48% (17), and misfitting raters comprised 13.25% (11).

Table 5 shows the distribution of raters according to consistency and whether a rater was an in-service teacher or a pre-service teacher. Among in-service teachers, fitting raters comprised 70.83%; misfitting raters, 16.67%; overfitting raters, 12.5%. Among pre-service teachers, fitting raters accounted for 60%; overfitting raters, 31.43%; and misfitting raters, 8.57%. Not only was the proportion of the fitting in-service teachers higher than that of pre-service teachers, but the proportion of misfitting in-service teachers was also higher. Further, the proportion of overfitting pre-service teachers was higher than the proportion of overfitting in-service teachers.

These results suggest that it is difficult for pre-service teachers to apply each rating scale evenly. In other words, in-service teachers use the full 6-point scales while pre-service teachers show a central tendency. In-service teachers build up their own point of view for assessing writing based on their writing instruction and assessment, whereas pre-service teachers assess writing samples depending only on the criteria or prompts. They may be familiar with general explanatory or argumentative essays, but not the writing task in the present study. This suggests that teachers as raters should be trained to assess writing in various genres in the same way that student should write in various genres.

Table 5. Rater Distribution According to Consistency and Roles

	In-service teachers(1-48)		Pre-service teachers(49-83)	
	Rater Number	Total(%)	Rater Number	Total(%)
fit	1,3,4,5,6,7,8,10,11,12,13,14,15, 16,17,18,20,21,22,23,25,26,27,2 9,30,32,33,37,40,41,42,43,44,46	34 (70.83%)	52,54,55,56,57,58,59,61,62, 63,65,67,68,70,71,74,76,78, 79,82,83	21 (60%)
over-fit	19,24,28,38,39,47	6 (12.50%)	50,51,53,60,64,66,72,73,75, 77,80	11 (31.43%)
mis-fit	2,9,31,34,35,36,45,48	8 (16.67%)	49,69,81	3 (8.57%)

Among the 27 male raters, 19 were fitting raters (70.37%), 5 were overfitting raters (18.52%), and 3 were misfitting raters (11.11%). Among the total of 56 female raters, 36 were fitting (64.29%), 12 were overfitting raters (21.43%), and 8 were misfitting raters (14.29%). Both male and female raters were displayed in order of whether they were fitting, overfitting, or misfitting.

2. Raters' writing assessment experience

As shown in table 6, a high degree of career separation reliability (.97) implies that the teachers could be reliably distinguished accord-

ing to their career length. As indicated by the chi-square value, the difference of career length was statistically significant.

Table 6. Career Measurement Report (arranged by MN)

Career	Obsvd	Fair(M)		Model	Infit		Outfit		Correlation	
	Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	PtMea	PtExp
3	3.91	3.77	-0.12	0.02	1.02	0.7	1.02	1.1	0.65	0.64
2	3.74	3.68	-0.02	0.02	1.07	2.5	1.10	3.5	0.65	0.65
1	3.61	3.63	0.02	0.01	0.89	-6.5	0.90	-5.9	0.64	0.65
4	3.50	3.54	0.12	0.02	1.15	5.5	1.15	5.4	0.64	0.62
RMSE .02 Adj (True) S.D. 0.10				Separation 5.27			Strata 7.36		Reliability 0.97	
Fixed (all same) chi-square: 82.1				d.f.: 3 significance (probability): 0.00						

* Career 1: pre-service teacher, 2: under 5 years, 3: 5-10 years, 4: 11-20 years

In-service teachers with less than 10 years of teaching scored more leniently, however in-service teachers with over 11 years of teaching experience scored more severely than pre-service teachers. This is similar to the findings of Park and Choi (2009), who wrote that the oldest group of teachers (with over 20 years of experience) and the youngest group of teachers (with less than 5 years of experience) scored lower than other groups that took part in the study. If teachers held a firm perception of writing, they might assess student essays more severely. This suggests that they might score on a lower rating scale than they would generally use. Changes in severity and the effects of using rating scales comparable to teaching experience should be studied more thoroughly.

Criteria are a critical factor for rater consistency, and reflect teachers' awareness of what constitutes a "good essay". According to the Criteria Measurement Report, the criteria were distinguished clearly with separation of 5.70 and reliability of 0.97; *Organization* was scored severely, and *Word choice* was scored leniently.

Table 7. Criteria Measurement Report (arranged by MN)

Criteria	Obsvd	Fair(M)		Model	Infit		Outfit		Correlation	
	Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	PtMea	PtExp
wor	3.81	3.79	-0.14	0.02	0.86	-6.1	0.87	-5.7	0.68	0.64
con	3.73	3.70	-0.04	0.02	1.05	1.8	1.06	2.2	0.66	0.65
exp	3.72	3.69	-0.03	0.02	1.04	1.5	1.05	1.8	0.63	0.65
sty	3.62	3.58	0.08	0.02	1.05	2.1	1.06	2.5	0.58	0.65
org	3.56	3.52	0.14	0.02	0.99	-0.4	1.00	0.1	0.68	0.65
RMSE .02 Adj (True) S.D. 0.11				Separation 5.70			Strata 7.93		Reliability 0.97	
Fixed (all same) chi-square: 10.3				d.f.: 1 significance (probability): 0.90						

Park and Choi (2010) reported that Korean language teachers scored *Organization, Expression* severely and *Style and mechanics, Word choice* leniently on explanatory writing assessment, while Choi and Park (2011) reported that Korean language teachers scored *Content, Style and mechanics* severely and *Organization, Word choice, Expression* leniently on argumentative writing assessment. Compared with previous study, it is similar that raters scored *Word choice* leniently. In present study, raters scored *Organization* severely. Because, despite of expressive writing, "interesting and original" and "own voice" (from table 2) related to *Expression* were difficult to score. Further, writings were written by imaging the situation according to the writing task, which is sort of letter, therefore raters focus on coherent and cohesive development.

A bias-interaction analysis can be performed to determine whether length of teaching career causes teachers to grade more severely or more leniently. There was no statistically significant bias with in-service or pre-service teachers and the criteria. Criteria scores were however reversed; the score of *Contents* and *Organization* is similar, though *Word choice* and *Style and mechanics* were very different.

For further analysis, a bias analysis for career length and criteria was also conducted. This is shown in table 8. Among 20 interactions,

4 were significant.

Table 8. Career × Criteria Bias Interaction Report

Career	Criteria	Bias Size	Model S.E.	T	Probability
2	exp	0.15	0.05	3.27	0.00
3	wor	0.09	0.04	2.30	0.02
3	exp	-0.11	0.04	-2.78	0.01
2	sty	-0.12	0.05	-2.67	0.01
Mean (Count: 10)		0.00	0.04	0.00	1.00
SD.		0.06	0.01	1.07	0.10
Fixed (all = 0) chi-square: 39.2 d.f.: 20 significance (probability): 0.01					

* Career 1: pre-service teacher, 2: under 5 years, 3: 5-10 years, 4: 11-20 years

The bias of teachers with less than 5 years and with 5~10 years of teaching experience were statistically significant. Based on career and 5 criteria, *Expression* showed the largest bias size and *Content* showed the smallest bias size. Teachers with less than 5 years of teaching experience tended to rate *Expression* more leniently than expected (Bias Size=0.15, t=3.27) and *Style and mechanics* more severely than expected (Bias Size=-0.12, t=-2.67). In contrast, teachers with 5~10 years of teaching experience were lenient about *Word choice* (Bias Size=0.09, t=2.3) and more severe about *Expression* (Bias Size=-0.11, t=-2.78). The evidence of score of *Style and mechanics* is visible and therefore both pre-service and novice teachers scored it severely. Experienced teachers, however, scored *Expression* according to the genre of the writing task.

3. Raters' gender

In boys' and girls' writing, there are differences by gender, topic, character, style, text, length, and so on. It is more difficult to find gender differences in essays written by university students (Francis, Read,

& Melling, 2003). This pattern of gender differences might be weakened as students advance through their education (Scheuer et al., 2011). This leads to some further questions: Do differences emerge if a particular rater rates an essay written by a male or a female? Is there any correlation with the gender of raters and the gender of students whose writing samples were rated?

Bias analysis with the gender of rater and examinee was also conducted. Table 9 shows that there were 4 significant biases among 8 biases.

Table 9. Examinee Gender × Rater Gender Bias Interaction Report

Examinee gender	Rater gender	Bias Size	Model S.E.	t	Probability
f	Pre-M	0.08	0.03	2.43	0.02
m	Pre-F	0.05	0.02	2.31	0.02
f	In-M	0.03	0.03	0.90	0.37
m	In-F	0.00	0.02	0.05	0.96
f	In-F	0.00	0.02	-0.04	0.97
m	In-M	-0.02	0.03	-0.87	0.38
f	Pre-F	-0.05	0.02	-2.36	0.02
m	Pre-M	-0.08	0.03	-2.36	0.02
Mean (Count: 8)		0.00	0.03	0.01	1.00
S.D.		0.05	0.01	1.85	0.10
Fixed (all = 0) chi-square: 24.0 d.f.: 8 significance (probability): .00					

Statistically significant bias emerged only by pre-service teachers; biases of male teacher were higher than female teacher. Male pre-service teachers rated girls' writing more leniently (Bias Size=0.08, t=2.43) while they rates boys' writing more severely (Bias Size=-0.08, t=-2.36). Also female pre-service teachers rated boy' writing more leniently (Bias Size=0.05, t=2.31) while they rates girls' writing more severely (Bias Size=-0.05, t=-2.36).

Although raters were offered no student information, they scored

more severely when they assessed an essay written by a student of the same gender. This supports the theory of *same-sex depreciation*. It seems that genre affects this phenomenon in writing assessment since Park and Choi (2009, 2010) reported no same-sex depreciation among teachers who graded explanatory or argumentative essays. The feature of bias between rater and student gender may have been revealed because an empathizing task was employed, thereby triggering an effect of genre. These were significant to pre-service teacher not in-service teacher. In-service teacher have experience of designing writing task, instruction and rating. Therefore they could score objectively than pre-service teachers. It seems that more experience of Korean teacher, less difference or bias of writing assessment. It suggest that gender bias of writing assessment had decreased by rating experience including interaction with co-worker and student

V. Summary and Conclusion

This study analyzed differences based on raters' career length and gender in writing assessment using an empathizing writing task. A summary of this study follows.

First, Korean teachers' rating severity is distinguished by their teaching experience and gender. In-service teacher rate student writing more leniently and this is supported by previous research. Among pre-service teachers, the proportion of raters who rate overfit consistency is higher than in-service teachers. This suggests that pre-service teachers hesitate to use the highest and lowest scores. On the other hand, when compared to pre-service teachers, not only do some teachers showed fit consistency, but also some teachers showed misfit consistency. The score may reflect various points of views held by teachers about the writing and empathizing task.

Second, scoring patterns were different based on teachers' career lengths and writing criteria. No bias was shown among pre-service

teachers and in-service teachers with more than 11 years of teaching experience. However, in-service teachers with less than 5 years of teaching experience scored the *Style and mechanics* criterion low, and those with 5~10 years of teaching experience scored the *Expression* criterion low. This implies that the scoring varied with each rater's understanding and application of the criteria.

Third, except in-service teachers, pre-service teachers had biases of gender significantly. Male raters graded girls' writing more leniently than they graded boys' writing. This supports the theory of *same-sex depreciation*, as it relates the grading of writing tasks to gender. This also shows that the genre or topic of the writing task can affect raters' grading of essays based on student gender and their own gender, and whether the essay is personal or emotional rather than explanatory or argumentative. In this situation, rater gender can also affect writing assessment. Therefore, when gender is examined, teachers should set criteria to prevent any effect of gender on their grading. The administrator of any large-scale writing assessment should control the ratio of rater gender.

These results also suggest that teachers should examine and reflect on their own writing assessment practices even if they have sufficient teaching experience. Therefore, "teachers as raters" need to be examined and could potentially revise their performance of writing assessments through specialized rater training.

* Submitted: 2014.10.31.
first Revision Received: 2014.12.06.
Accepted: 2014.12.06.

REFERENCES

Bouwer, R., Béguin, A., & Sanders, T. (2014). Effect of genre on the generalizability of writing scores, *Language Testing*, 4, 1-18.

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

Etaugh, C., Houtler, B. D., & Ptasnik, P. (1988). Evaluating competence of women and men, *Psychology of Women Quarterly*, 12(2), 191-200.

Francis, B., Read, B., & Melling, L. (2003). University lecturers' perceptions of gender and undergraduate writing, *British Journal of Sociology of Education*, 24(3), 357-373.

Harari, H., & McDavid, J. W. (1973). Name stereotypes and teachers' expectations, *Journal of Educational Psychology*, 65(2), 222-225.

Haswell, R. H., & Haswell, J. T. (1996). Gender bias and critique of student writing, *Assessing writing*, 3(1), 31-83.

Klein, K. J., & Hodges, S. D. (2001). Gender differences, motivation, and empathic accuracy: When it pays to understand, *Personality and Social Psychology Bulletin*, 27(6), 720-730.

Linacre, J. M. (1989). *Many-faceted Rasch Measurement*. Chicago: MESA Press.

Mertler, C. A. (2005). Secondary teachers' assessment literacy: Does classroom experience make a difference?, *American secondary education*, 33(2), 49-64.

Peterson, S. (2000). Fourth, sixth, and eighth graders' preferred writing topics and identification of gender markers in stories, *The Elementary School Journal*, 101(1), 79-100.

Peterson S., & Kennedy, K. (2006). Sixth grade teachers' written comments on student writing: Genre and gender influences, *Written Communication*, 23(1), 36-62.

Roen, D. G. (1992). Gender and teacher response to student writing, In McCracken, N. M., & Appleby, B. C. (eds.). *Gender Issues in the Teaching of English* (pp. 126-141). Portsmouth, NH: Boynton.

Scheuer, N., de la Cruz, M., Pedrazzini, A., Iparraguirre, M. S., & Pozo, J. I. (2011). Children's gendered ways of talking about learning to write, *Journal of Writing Research*, 3(3), 181-216.

Schulte-Rüther, M., Markowitsch, H. J., Shah, N. J., Fink, G. R., & Piefke, M. (2008). Gender differences in brain networks supporting empathy, *Neuroimage*, 42(1), 393-403.

Toussaint, L., & Webb, J. R. (2005). Gender differences in the relationship between

empathy and forgiveness, *The Journal of social psychology*, 145(6), 673-685.
Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making, *Assessing Writing*, 17(3), 150-173.
Whithaus, C., Harrison, S. B., & Midyette, J. (2008). Keyboarding compared with handwriting on a high-stakes writing assessment, *Assessing Writing*, 13(1), 4-25.
박영민(2010),「예비국어교사의 쓰기 평가 효능감 분석」,『청람어문교육』42, 181-207.
_____ (2011),「국어교사의 쓰기 평가 효능감 분석」,『청람어문교육』44, 121-146.
_____ (2013),「현직 국어교사와 예비 국어교사의 쓰기평가지식의 차이 분석」,『작문연구』19, 331-352.
박영민·최숙기(2009),「현직 국어교사와 예비 국어교사의 쓰기 평가 비교 연구」,『교육과정평가연구』12(1), 123-143.
박영민·최숙기(2010),「예비 국어교사의 성별 및 학생 성별 인식에 따른 평가 차이 분석」,『교육과정평가연구』13(2), 239-258.
박종임(2013),「국어교사의 쓰기평가 특성 연구」, 박사학위논문, 한국교원대학교.
박종임·박영민(2011),「Rasch 모형을 활용한 국어교사의 채점 일관성 변화 양상 및 원인 분석」,『우리어문연구』39, 301-335.
배영태(2010),「반성적 쓰기가 고등학생의 수필 쓰기 능력 및 태도에 미치는 영향」, 한국교원대학교 석사학위논문.
법정(1999),『무소유』, 범우사.
정미경(2011),「국어교사의 성별에 따른 쓰기평가 특성 분석」,『교원교육』27, 73-93.
최숙기·박영민(2011),「논설문 평가에 나타난 국어교사의 평가 특성 및 편향 분석」,『교육과정평가연구』14(1), 201-228.

ABSTRACT

Comparisons of In-service and Pre-service Korean Language Teachers' Writing Assessments Using an Empathizing Task

Jang, Eun Ju · Park, Young Min

The purpose of this study was to examine writing assessment according to raters' career length and gender. A total of 83 teachers were asked to rate essays written by students following completion guidelines in an empathizing task. The results showed that writing assessments differed by teachers' career length and by assessment criteria, and that assessments changed based on the gender of both the students and raters, thus revealing gender bias and same-sex depreciation. These outcomes suggest the following: First, teachers should control writing task criteria. Second, rater gender should be controlled in large-scale writing assessments. Lastly, teachers should examine and reflect on their own writing assessment skills even if they have sufficient teaching experience.

KEYWORDS writing assessment, empathic essay writing, rater bias, rater's gender, pre-service Korean language teachers, in-service Korean language teachers