영어어문교육 16권 1호 2010년 봄

A FACETS Analysis of Rater Characteristics and Rater Bias in Measuring L2 Writing Performance

Yousun Shin
(Pukyong National University)

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.

The present study used multi-faceted Rasch measurement to explore the characteristics and bias patterns of non-native raters when they scored L2 writing tasks. Three raters scored 254 writing tasks written by Korean university students on two topics adapted from the TOEFL Test of Written English (TWE). The written products were assessed using a five-category rating scale (Content, Organization, Language in Use, Grammar, and Mechanics). The raters only showed a difference in severity with regard to rating categories but not in task types. Overall, the raters scored Grammar most harshly and Organization most leniently. The results also indicated several bias patterns of ratings with regard to the rating categories and task types. In rater-task bias interactions, each rater showed recurring bias patterns in their rating between two writing tasks. Analysis of rater-category bias interaction showed that the three raters revealed biased patterns across all the rating categories though they were relatively consistent in their rating. The study has implications for the importance of rater training and task selection in L2 writing assessment.

[bias analysis/L2 writing tasks/task types/rater variability/rater training effects]

I. INTRODUCTION

As interest has grown in performance-based language assessment, the variability of ratings has been a major issue (Jones, 1985; McNamara, 1996; Shohamy, Gordon, & Kraemer, 1992; Song, 2002). Studies of rater characteristics in performance-based assessment have pointed out that variability among raters considerably affected learners' spoken or written performance (Bachman, Lynch, & Mason, 1995; Eckes, 2005; Weigle, 1998). Since rater variability is not linked to the performance of learners but to the characteristics of the raters, it inevitably weakens the reliability of an assessment battery.

Researchers have agreed that as a complex and subjective process, rating of written products leads to systematic variance so that the effect of the rater needs to be examined (Myford & Wolfe, 2003; Schaefer, 2008).

Many researchers have investigated rater behavior using multi-faceted Rasch measurement (Eckes, 2005; Kondo-Brown, 2002; Lynch & McNamara, 1998; Schaefer, 2008). Bias analysis using a multi-faceted Rasch measurement examines rater variability in terms of other facets in the Rasch model. The term "bias" refers to rater severity or leniency in scoring, whereas bias analysis identifies systematic subpatterns of rating behavior from an interaction of a rater and some other aspect of the rating situation (McNamara, 1996). For example, if a particular rater is unusually harsh in his or her ratings, this would be taken into consideration when estimating the ability of a learner by that rater. It is expected that bias analysis guides researchers to discover the sources of rater bias, which, in turn, helps to systemize rater training and facilitate rating criteria refinement which best fits into an EFL context like Korea.

The present study aims to explore if raters evaluate certain aspects of rating categories and certain writing tasks harshly or leniently. Previous studies that examined biased rater characteristics (Lynch & McNamara, 1998; Schaefer, 2008) extensively analyzed the interaction between raters, learners, and tasks with regard to the effects of rater training. The present study attempts to examine bias patterns of non-native raters using the FACETS program. The program was run to identify the factors that contribute to rater severity or leniency when assessing L2 writing ability. More specifically, which features in L2 writing assessment, as operationalized in rating categories, would affect raters' severity or leniency. It also explored if raters behaved differently in the scoring of written products depending on the task types. Accordingly, researchers in Korea will be more cognizant of the characteristics of non-native raters which would be unquestionably different from these of native raters so that the results will lead to better understanding of non-native raters and benefit to devise a more suitable rater training programme in Korea.

II. LITERATURE REVIEW

A number of studies have examined unexpected interactions between raters' judgments and learners' performance or other facets of bias analysis. For example, Wigglesworth (1993) examined rater-item, rater-task, and rater-test type interactions when assessing speaking ability to determine if rater training improved rating behavior. Raters displayed individual patterns when using scoring criteria. Some raters were consistent in their rating while others were not. Moreover, the raters differed from one another in their harshness or leniency with respect to the different task types. The results of the bias analysis were

provided to the raters as feedback. Subsequently, the feedback was considered effective because raters, in turn, demonstrated some improvement in their rating consistency during the second round of rating. In her study, bias analysis in rater training proved to be a very useful tool for building a profile of rater characteristics and obtaining continuing feedback on a rater's individual performance.

McNamara (1996) analyzed results of the Occupational English Test and concluded that experienced raters were affected by learners' grammatical accuracy, which contradicted the communicative nature of the test. The analysis revealed that there was a significant mean difference between a rater's perception of the importance of grammatical accuracy and the actual results of the test. That is, experienced raters did not regard grammatical accuracy as a priority; however, the results revealed grammar proved to be important as one of the rating features. He further noted that the multi-facet Rasch model uncovered these underlying patterns in the ratings.

Similarly, Weigle (1998) examined differences in rater severity and consistency among inexperienced and experienced raters before and after rater training using the FACETS program. A total of 16 raters scored 60 essays before and after rater training with three rating categories: content, rhetorical control, and language. The analysis showed that inexperienced raters tended to be both more severe and less consistent in their ratings than experienced raters, before training. The overall differences between the two groups of raters were not statistically significant. After rater training, significant differences in severity persisted while consistency improved. From her study, it seems apparent that rater training is more helpful with regard to rater consistency rather than with regard to rater severity.

Schaefer (2008) investigated rater bias patterns of native English-speaking raters. Forty native English-speakers rated 40 essays with six rating categories: content, organization, style and quality of expression, language use, mechanics, and fluency. In the analysis of rater-category bias interactions, content and organization tended to be rated severely while language use and mechanics tended to be rated leniently. In rater-writer bias interactions, higher ability writers were found to be rated more severely or leniently than lower ability writers.

With increasing attention to performance tests along with the validation process in Korea (i.e. Choi, 2005; Choi, 2001, 2002), Choi, (2001) investigated the effects of tasks, raters, and rating scales on test scores. 38 high school students were asked to complete two different writing tasks. Their written products were scored by two Korean raters and one American rater based upon the 9-point-scale analytic scoring criteria. They were analyzed by using GENOVA and FACETS. The result of FACETS indicated significant variations among the three raters, which might be caused by the American rater's rating behavior. She noted that the difference might have been grounded by unidentical training sessions among

two Korean raters and one American rater. The result of the analysis found out, though, that each rater was consistent in their ratings similar to the results of previous research (i.e. Kondo-Brown, 2002; Lumley & McNamara, 1995; McNamara, 1996; Weigle, 1998).

Most studies indicated that there are significant differences in rater severity while rater consistency can improve through rater training (i.e. Choi, 2002; McNamara, 1996; Weigle, 1998; Wigglesworth, 1993). Bias analysis allows us to gain an understanding of the nature of individual rater characteristics. Increased rater consistency through rater training and bias analysis contributes to reducing the error of measurement, and subsequently increases fairness and accuracy in performance-based assessments (Wigglesworth, 1993). The findings from numerous studies (Choi, 2001, 2002; Lumley & McNamara, 1995; McNamara, 1996; Weigle, 1998), thus, would be beneficial in improving the understanding of rater behavior and provide more detailed information in the nature of rater training.

Particularly in Korea, a variety of language performance tests has been gradually used to identify learners' language ability and improve their language skills (Rafik-Galea, 2003; Song, 2002). Under the circumstance, teachers, parents, or administrators have called for reliable and valid performance tests, while researchers have attempted to develop not only useful assessment tools but also appropriate rater training programs corresponding to each assessment tool. Thus, as a first step, Korean raters' rating patterns, if any, need to be identified in order to improve their rating behavior followed by proper training sessions. Consequently, the raters will be expected to yield reliable and valid scores on learners' performance through a series of the useful analysis and following rater training process.

III. METHODOLOGY

1. Participants

1) Learners

The learners who participated in the study were 127 nonnative speakers of English with the same Korean L1 background. The data used in this study were mainly collected from the students enrolled at three universities in Korea during the spring 2007 semester. Of the 127 participants, 54 were males and 73 were females. The participants varied from first-year to fourth-year university students and their majors were also diverse including majors in English, engineering, business, history, or accounting. Learners writing ability varied from novice to advanced judging from the analytic scores of the essays.

2) Raters

Three Korean Ph.D candidates, who majored in English language education and had been teaching English written communication courses in a university setting, rated all the learners' written products. The raters were all Koreans as non-native English speakers and they were all females. The rates had at least two years prior teaching experience and were quite familiar with the rating scales, the writing prompts, and the level of university learners' writing and their writing proficiency.

2. Writing Tasks

The participants were given one argumentative writing task and one expository writing task. They completed the two tasks over two-week periods. The participants had 10 minutes for planning and 30 minutes (i.e. Ellis & Yuan, 2004; Kellogg, 1988, 1990) to complete a task. The participants who finished both writing tasks were included in the analysis and did not have any specific instruction about how lengthy the essay should be. The topics of the two writing tasks were as follows:

1. Expository task: Movies or TV dramas can tell us about the country where they were made. What have you learned about a country from watching its movies or TV dramas? Use specific examples and reasons to support your ideas.
2. Argumentative task: Do you agree or disagree with the following statement? Children need to learn a foreign language as soon as they start school. Use specific reasons and examples to support your position.

Both tasks were adapted from the Test of Written English (TWE) sample writing topics available to the public (http://www.toefl.org). The expository writing task and the argumentative writing task were chosen in that both task types are not only considered to be typical of academic tasks (Grabe, 2001; Silva, 1993) but also expected to elicit different features of writing tasks.

3. Scoring Procedure

Each rater scored a whole set of 254 written products. A two-hour rater training session was held before the raters began to score, which was led by the researcher herself. In the training session, raters were provided a pre-set of instructions which defined the dimensions and short descriptions of the scoring rubric used in the study. Specifically, raters discussed the writing topics and the rating rubric and scale, read anchor essays that

represent the various scores on the scale, and then practiced rating essays as the final step of the rating session.

Following the instructions, raters were required to read quickly through all sets of essays before assigning a score to any composition. It is anticipated that the raters could catch a glimpse of overall writing proficiency of the participants through the process. Afterward, they had enough time to judge the essays accurately and fairly in a setting where they felt comfortable. Raters were asked to use full 5-point scales in the analytic scoring process. In order to avoid the effect of handwriting, the 254 handwritten essays had been word-typed verbatim with spelling or grammatical errors left intact and the original format of the writings also preserved as Uzawa (1996) recommended.

4. Rating Rubric

The quality of the learners' written products was evaluated by means of analytic rating scales (see APPENDIX for the rating scales). Analytic scales were used to yield multiple scores that separately assess more than one component of the text (Kroll, 1998). Analytic scoring allows raters to identify different aspects of writing so that each score identifies detailed diagnostic information about the learners' writing performance (Kroll, 1998). A modified version of Cohen's (1994) and Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey's (1981) scale was used for this study. It focused on five aspects of writing: Content, Organization, Language in Use, Grammar, and Mechanics. Analytic scoring is indicated as a band score ranging from 1 to 5.

IV. RESULTS AND DISCUSSION

The ratings were analyzed using FACETS, a multi-faceted Rasch measurement computer software program (Linacre, 2005). The data were categorized as four facets: the ability of the learners, the difficulty of the tasks, the difficulty of the rating categories, and the severity of the raters. The results were provided on a common scale, *logits*, as the basic unit of measurement in the analysis. In the analysis, the raters were ranked in relation to how likely they were to assign particular scores to particular learners on particular tasks.

1. FACETS Summary

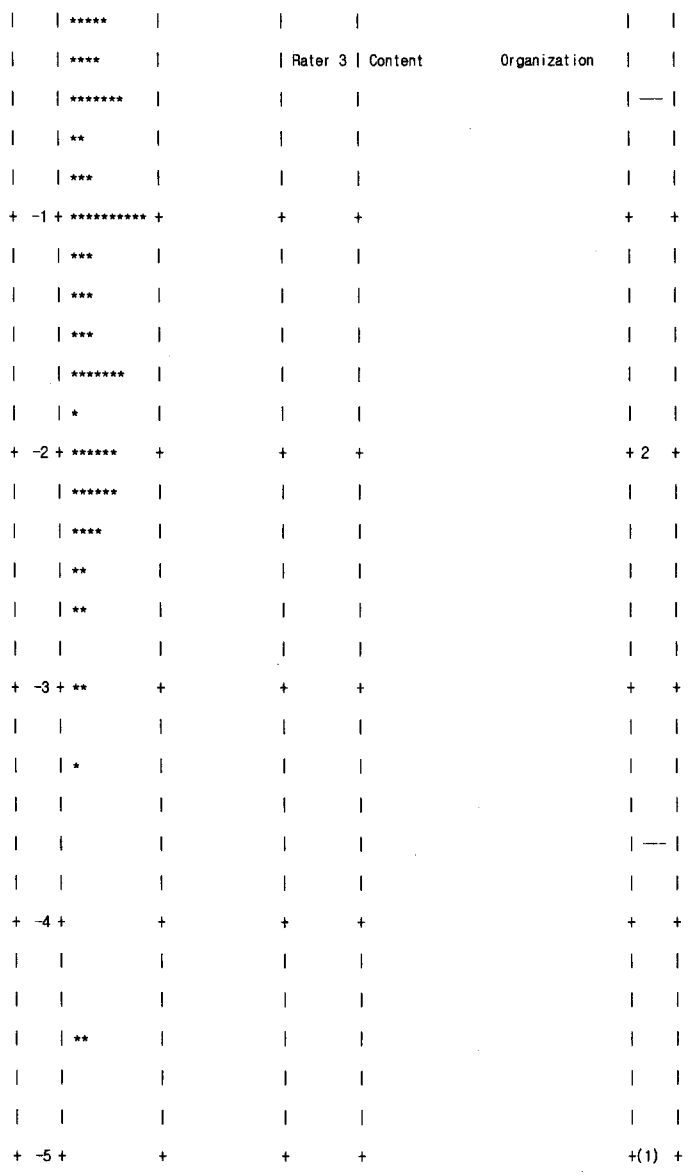
Figure 1 describes the measures for learner ability, task difficulty, rater severity, and rating category difficulty, respectively. The scale in the first column is the logit measure. The second column shows the ability variability among learners. The top of the scale

indicates higher ability, and the bottom of the scale indicates lower ability. Learners are spread out along the measure, with slightly less than half above 0.00 logits, but the majority of the learners are between -1.00 and +1.00 logits. The third column indicates task difficulty, showing no difference in task difficulty between the two writing tasks. The fourth column shows the severity variability among the raters. The most severe rater is Rater 2 and the least severe rater is Rater 3. The fifth column indicates difficulty variability among the rating categories. It is shown that all the rating categories clustered around the mean. The most severely rated category is Grammar while Content and Organization are perceived as the easiest categories to rate. The last column shows the rating scale measure of five points and the distance between the steps on the scale. It corresponds to the scores that the learners at any ability level on the scale are likely to receive (Weigle, 1998). For instance, a learner at a 1.00 logit ability level has a 50% probability of getting 3 points on any rating category from a rater at a 1.00 logit severity level.

FIGURE 1

FACETS Summary (learner ability, task difficulty, rater severity, category difficulty)

Logit	Learner	Task	Rater	Rating Category	Score
+ 3 +		+	+	+	+(5) +
| |		|	|	|	| |
| |	*	|	|	|	| |
| |		|	|	|	| |
| |		|	|	|	| |
| |		|	|	|	| |
| |		|	|	|	| |
+ 2 +	**	+	+	+	+ 4 +
| |	**	|	|	|	| |
| |	***	|	|	|	| |
| |	****	|	|	|	| |
| |	**	|	|	|	| — |
| |	****	|	|	|	| |
+ 1 +	***	+	+	+	+ +
| |	********	|	|	|	| |
| |	*****	|	|	|	| |
| |	*******	|	|	|	| 3 |
| |	********	|	| Rater 2 |	Grammar	| |
| |	**	|	| Rater 1 |	Mechanics	| |
* 0 *	***	* task1 task2 *		*	Language in use	* *



Notes: The horizontal dashed lines in the sixth column indicate the rating category threshold measures for the five-point rating scale. The asterisk (*) indicates one learner.

Figure 1 shows that the number of learners below 0 logit is more than the number of learners above 0 logit, which means that the learners' writing ability is relatively lower with regard to the level of task difficulty. This mismatch between learners ability and task difficulty can be a significant issue depending on the purpose of the test (McNamara, 1996).

It is possible that the purpose of the writing tasks in the study was simply to identify their L2 writing ability. Accordingly, learners might have felt burdened by the two timed essay tasks. That might have partially affected their performance to measure their L2 writing skills. It is also plausible that since the results of the two writing tasks were not to be included in their final grade, they might have not done their best to complete the given tasks.

2. Difficulty level for two task types and five rating categories

Table 1 summarizes the results of the FACETS analysis for task type. The result indicated that there was not significant mean difference between the task types. It implies that the two tasks were equivalent in difficulty.

TABLE 1

Difficulty Measurement Report for Task Types

Observed scores	Observed count	Observed Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	No.	Task type
5098	1905	2.7	2.57	-.03	.03	.9	-2	.9	-2	2	Task2
5042	1905	2.6	2.54	.03	.03	1.1	1	1.0	1	1	Task1
5070.0	1905.0	2.7	2.55	.00	.03	1.0	-.5	1.0	-.4		Mean(Count2)
28.0	.0	.0	.01	.03	.00	.1	2.2	.1	1.8		S.D.

RMSE (Model).03 Adj S.D. .00 Separation .00 Reliability .00

Fixed (all same) chi-square: 1.8 df:1 significance: $p < .18$

Infit statistics demonstrate that the scores of the two task types are predictable, which means that the rating of the two tasks were consistent, as shown in Table 1. A value which is greater than the mean plus twice the standard deviation of Infit Mean Square would be considered as misfitting data (McNamara, 1996), that is, there is too much unpredictability in rating. In this case, mean of the Infit Mean Square (1.0) plus twice the standard deviation of it (2×0.1), a value more than 1.2 would be misfitting. Thus, none of the tasks is misfit and the two tasks demonstrated similarity in difficulty level. Table 2 provides a detailed difficulty measurement with regard to the five rating categories.

TABLE 2

Difficulty Measurement Report for Rating Categories

Observed scores	Observed count	Observed Average	Fair-M Average	Model Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	No.	Rating Category
2157	762	2.8	2.72	-.37	.05	1.1	1	1.1	1	2	Organization
2122	762	2.8	2.67	-.27	.05	1.2	3	1.2	3	1	Content
2000	762	2.6	2.52	.07	.05	.9	-2	.9	-2	3	Language in Use
1970	762	2.6	2.48	.16	.05	.9	-2	.9	-1	5	Mechanics
1891	762	2.5	2.39	.40	.06	.9	-2	.8	-3	4	Grammar
2028.0	762.0	2.7	2.56	.00	.05	1.0	-.5	1.0	-.3		Mean(count:5)
98.4	.0	.1	.12	.28	.00	.1	2.8	.1	2.6		S.D.

RMSE (Model) .05 Adj S.D. .28 Separation 5.18 Reliability .96

Fixed (all same) chi-square: 139.1 df: 4 significance <.00

Table 2 indicates that the most harshly scored category was Grammar while the most leniently scored category was Organization. The high separation index (5.18), the high reliability index (0.96), and the chi-square of 139.1 (df = 4) demonstrated that significant variation in difficulty existed among the five rating categories. In other words, raters consistently scored grammar (0.40 logits) more harshly than other four categories (Organization (-0.37 logits), Content (-0.27 logits), Language in Use (0.07 logits), Mechanics (0.16 logits)).

With regard to Infit statistics, mean of the Infit Mean Square mean (1.0) plus twice the standard deviation (2 × 0.1), a value more than 1.2 would be misfitting which is considered as unpredictable scores. Thus, Infit statistics of the rating categories indicated that only one category, Content (1.2), was identified as a borderline for misfitting. According to Lynch & McNamara (1996), the type of inconsistency, in this case, occurred from scoring Content, cannot be modeled and compensated for by the program so it should be improved by further training sessions.

3. Rater Characteristics

1) Rater Severity

A detailed analysis of rater behavior is shown in Table 3 including the raters' measurement report for learners' written products.

TABLE 3

Rater Measurement Report of Raters

Observed scores	Observed count	Observed Average	Fair-M Average	Model Measure	Model S.E.	Infit MsSq	Infit ZStd	Outfit MnSq	Outfit ZStd	No.	Rating Category
3622	1270	2.9	2.74	-.41	.04	.8	-5	.8	-5	3	Rater3
3313	1270	2.6	2.51	.11	.04	1.2	4	1.2	4	1	Rater1
2305	1270	2.5	2.43	.30	.04	1.0	0	1.0	0	2	Rater2
3380.0	1270.0	2.7	2.56	.00	.04	1.0	-.5	1.0	-.5		Mean(count:3)
176.7	.0	.1	13	.30	.00	.2	4.1	.2	4.1		S.D.

RMSE (Model) .04 Adj S.D. .30 Separation 7.18 Reliability .98
Fixed (all same) chi-square: 149.9 df : 2 significance: $p < .00$

Raters are presented in descending order of severity; in other words, Rater 2 (0.30 logits) is more severe than the others whereas Rater 3 (-0.41 logits) is more lenient than the others. With high separation index (7.18) and high reliability index (0.98), the chi-square of 149.9 ($df = 2$) was statistically significant at $p < .00$. Hence, it can be concluded that there was significant variation in severity among the three raters. Infit statistics show that none of the raters unpredictably scored, which means that their ratings were consistent, as presented in Table 3. In this case, a value more than 1.4 (Mean of the Infit Mean Square (1.0) + twice the standard deviation (2×0.2)) would be misfitting.

1) Bias Analysis

Bias analysis, as it is called by FACETS, allows us to investigate pairs of facets and report the number of inconsistent biased patterns of responses from a rater. These analyses provide individual raters with information including their relative characteristics as raters, their consistency, and any biased ratings (McNamara, 1996). A bias analysis of the interaction between the facets of a rater and a task, and between the facets of a rater and rating categories can provide information about whether each rater was consistently harsh or lenient on any particular aspect (McNamara, 1996).

(1) Rater-task Bias Interactions

A bias analysis was also carried out on the interaction of raters with the two tasks. The meaning of bias here is that z-scores greater than ±2 are considered to be significantly biased, given rater bias always exists to some extent (McNamara, 1996).

TABLE 4

Bias Calibration Report: Rater-Task Interactions

Rater	Task	Rater severity (logit)	Rating category (logit)	Observed scores	Expected scores	Bias (logit)	Error	z-score
Rater1	Task2	-.03	.11	1787	1665.7	.19	.06	-7.17
Rater2	Task1	.03	.30	1687	1593.7	.15	.06	-5.65
Rater3	Task1	.03	.41	1829	1801.0	-.09	.06	-1.60
Rater3	Task2	-.03	-.41	1793	1821.0	.09	.06	1.59
Rater2	Task2	-.03	.30	1518	1611.4	-.15	.06	5.64
Rater1	Task1	.03	.11	1526	1647.4	-.19	.06	7.22

Notes: Fixed (all = 0) chi-square: 172.4 df: 6 significance: $p < .00$

Table 4 shows the results of rater-task interaction. Four out of six possible interactions (three raters × two tasks) showed significant bias interactions. This suggests that the two writing tasks were not effectively measured by the raters. The analysis identified Rater 1, and 2 as biased. The output also indicated both the direction and the extent of the bias. Two were negative bias z-scores (showing leniency) and two were positive bias z-scores (showing severity). Rater 1 was harsher in rating Task 1 and more lenient in rating Task 2. On the contrary, Rater 2 shows a reverse of the pattern observed from Rater 1. Rater 2 was more lenient in rating Task 1 and more severe in rating Task 2. Thus, the result summarizes that Rater 1 and Rater 2 showed significant changes in their severity between Task 1 and Task 2.

FIGURE 2

Rater-Task Bias Interactions by Tasks

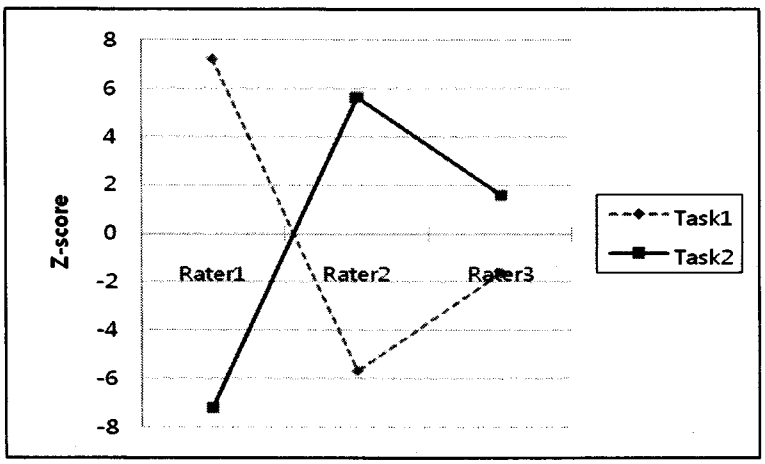


Figure 2 indicates that there were bias patterns specific to each task. Rater 1 is more severe on the rating of Task 1 and more lenient on the rating of Task 2. On the other hand,

Rater 2 is more severe on the rating of Task 2 and more lenient on the rating of Task 1. Thus, it is assumed that the graph has an X shape due to the opposite direction of the two raters in rating each task. Last, Rater 3 seemed to be more severe on the rating of Task 2 and more lenient on the rating of Task 1 but the rating differences were not as notable as the other raters. In sum, Rater 1 and Rater 2 appear to be inconsistent with respect to harshness and demonstrate high variability in task types. Rater 3 demonstrates, on the contrary, fairly consistent behavior in both task types in terms of the harshness of rating.

(2) Rater-category Bias Interactions

Table 5 shows the results of the bias analysis of the rater-rating categories interactions. It lists 12 interactions with a significant bias out of the total 15 interactions (three raters × five rating categories).

TABLE 5

Bias Calibration Report: Rater-Category Interactions

Rater	Rating category	Rater severity (logit)	Rating category (logit)	Observed scores	Expected scores	Bias (logit)	Error	z-score
Rater1	Content	.11	-.27	790	693.4	-.77	.09	-8.80
Rater3	Language in use	-.41	.07	768	714.5	-.43	.09	-4.82
Rater3	Grammar	-.41	.40	716	675.5	-.34	.09	-3.76
Rater1	Organization	.11	-.37	746	704.9	-.33	.09	-3.74
Rater2	Mechanics	.30	.16	659	622.6	-.33	.09	-3.53
Rater2	Grammar	.30	.40	623	597.8	-.24	.10	-2.50
Rater2	Organization	.30	-.37	677	681.9	.04	.09	.45
Rater2	Language in use	.30	.07	625	632.1	.07	.10	.68
Rater3	Mechanics	-.41	.16	693	703.8	.09	.09	.98
Rater1	Mechanics	.11	.16	618	643.6	.24	.10	2.44
Rater3	Organization	-.41	-.37	734	70.2	.29	.09	3.19
Rater3	Content	-.41	-.27	711	757.9	.38	.09	4.14
Rater1	Language in use	.11	.07	607	653.4	.43	.10	4.38
Rater2	Content	.30	-.27	621	670.7	.45	.10	4.63
Rater1	Grammar	.11	.40	552	617.8	.67	.10	6.41

Notes: RMSE (Model) .05 Adj S.D. .28 Separation 5.18 Reliability .96

Fixed (all = 0) chi-square: 264.2 df: 15 significance: $p < .00$

As described before, a z-score below -2.0 as a bias estimate indicates that the rater consistently scored the category more leniently than the model predicted. Conversely, a z-

score above +2.0 indicates that a rater consistently scored the category more harshly than expected. The table 5 also illustrates that significant bias interactions were equally distributed among the raters (five interactions for Rater 1, three interactions for Rater 2, and four interactions for Rater 3). There were three significant bias interactions for Content, two for Organization, two for Language in use, three for Grammar, and two for Mechanics. The raters were likely to show severe or lenient bias towards Grammar and Content. Figure 2 graphically illustrates the information on rater-rating category interactions based on the distribution of the z-scores.

FIGURE 3

Rater-Category Bias Interactions by Raters

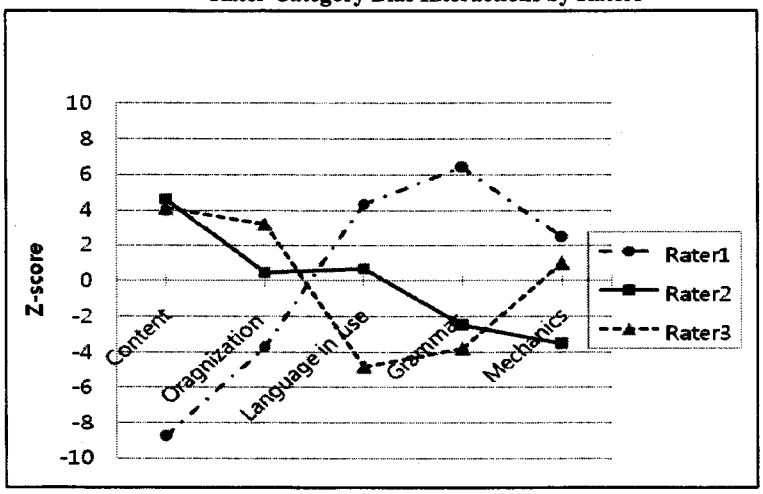


Figure 3 indicates that there were bias patterns specific to each rater. Rater 1 is more severe on the rating of grammar and more lenient on the rating of content. Rater 2 is more severe on the rating of content and more lenient on the rating of mechanics. Last, Rater 3 is more severe on the rating of Content and more lenient on the rating of Language in Use.

The tendency toward leniency on Grammar demonstrated by Rater 2 and Rater 3 is unusual, though. Previous research has claimed that raters generally score grammatical structure more harshly than other aspects of learners' performance (Lumley & McNamara, 1995; Wigglesworth, 1993). Moreover, a number of studies have raised the issue of rater characteristics based on the distinction between nonnative speakers and native speakers (i.e., Kondo-Brown, 2002; Shi, 2001). Non-native raters tend to rate grammar more harshly than other categories (Eckes, 2008; Kondo-Brown, 2002; Schaefer, 2008). Native English raters are inclined to attend more positively to content and language, whereas nonnative raters usually place more emphasis on the general organization and length of the essay (Shi, 2001). As one of the reasons, native English teachers may have instructional focuses that

differ from nonnative English teachers in their writing classes (Mohan & Lo, 1985; Shi, 2001) and these may also be reflected in their assessment of essays. It is also found out that inexperienced raters are likely to harshly score learners' written or spoken performance than experienced raters (Choi, 2002; Shi, 2001; Weigle, 1998).

In sum, Rater 2 and Rater 3's tendency toward leniency on Grammar could be speculated by the following reasons to some extent: (1) the raters might more focus on content and organization as their priorities so that they did not consciously pay attention to grammar in scoring learners' written products, (2) Rater 2 and Rater 3 might have had more scoring experiences than Rater 1, and (3) the scoring criteria and descriptors might be not clear enough for the raters to yield a consistent score in certain categories, particularly Content, Language in Use, and Grammar.

V. CONCLUSION

The present study attempted to examine if non-native raters rated certain aspects of rating categories and certain writing tasks harshly or leniently. The findings of the investigation into rater consistency were reported based on the data obtained from the two writing tasks. The FACETS analysis was used to investigate whether the non-native raters scored particular categories or particular task types more or less harshly than others. Unlike Choi's (2001) research results, the result of the analysis in this study demonstrated that the rating categories significantly affected the test scores rather than the task types. It might be caused by the fact that the present study employed two similar writing tasks so that the FACETS analysis did not provide enough information on the facet – task types. The results of the study are summarized that:

1. The raters only showed a difference in severity with regard to rating categories but not in task types: the raters scored Grammar most harshly and Organization most leniently overall;
2. The raters demonstrated biased patterns in task types though they were relatively consistent in their rating: Rater 1 was harsher in rating Task 1 and more lenient in rating Task 2 while Rater 2 showed a reverse of the pattern with Rater 1, that is, more lenient in Task 1 and harsher in Task 2.
3. The three raters evenly demonstrated biased patterns across all the rating categories – Content, Organization, Language in Use, Grammar, and Mechanics though they were relatively consistent in their rating.

The results have revealed that the three raters showed self-consistent scoring patterns

across the writing tasks, though each rater was biased towards certain type of tasks and rating categories. The findings of the present study were consistent with the results of previous studies on rater variability in that a rater's tendency for severity or leniency continues but highly self-consistent in their scoring (Choi, 2001; Kondo-Brown, 2002; Lumley & McNamara, 1995; McNamara, 1996; Weigle, 1998). In the meanwhile, the results of this study have pointed that it needs to develop effective rater training based on the raters' behaviors. The results suggest that raters will be able to identify their own rating problems and improve intra-and-inter-rater reliability to a certain extent by providing more systematic and recurring rater training sessions.

As assessment specialists attempt to measure writing ability in a variety of testing contexts with different testing purposes, more effective training and discussion sessions are needed to reach consensus on their judgments for yielding reliable scores on learners' writing ability. In this sense, the results of FACETS can be served as feedback about a rater's rating behavior in a particular task environment. Previous research results maintained that some raters tend to be more severe or more lenient than others, no matter how much training raters are provided (Lumley & McNamara, 1995; Weigle, 1998; Wigglesworth, 1993). Indeed, rater training can improve rater consistency in writing assessment (Choi, 2001; Lumley & McNamara, 1995; Weigle, 1998; Wigglesworth, 1993). Tools such as FACETS can be useful when incorporating the result of rater bias into a rater training program so that at least raters can be reminded of their own rating behavior and rater severity while improving rating consistency in their writing.

One of the limitations in the study is that only three Korean raters were employed in the rating process. Consequently, it is not possible to compare their rating behaviors to native raters' ones and the results from this study may not be generalizable outside of this specific context. However, the study has valuable implications for L2 writing assessment. First, the FACETS analysis used in the study enabled us to investigate differences in rater behavior in terms of severity and consistency across different writing tasks. Traditional methods comparing mean scores are not able to examine these types of the differences in rater behavior. Second, as FACETS can be a useful tool in taking rater severity/leniency into account, score users including decision makers, language professionals or parents get informed about to what extent raters differ from each other in terms of severity/leniency or rater consistency (Weigle, 2000).

Further research needs to compare non-native rates' rating pattern with native raters' one in combination with the effect of training in order to determine (1) whether similar or different patterns would emerge from a similar group of raters and (2) how and to what extent training sessions help raters to minimize each rater's different perceptions on a scoring criteria and to reach satisfactory consensus on learners' writing or spoken performance. Additionally, rater-training time effects need to be further tested to see if rater

training actually improves the validity of an assessment over a certain period of time and how long the effect of training session will be sustained.

REFERENCES

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257.

Choi, I. C. (2005). ASR-based simulated oral proficiency interview model for performance for middle school English. *Foreign Languages Education, 12*(4), 235-266.

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

Choi, Y. H. (2001). GENOVA and FACETS analysis of an English writing test: Tasks, raters, and scales. *English Teaching, 56*(2), 125-142.

Cohen, A. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221.

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

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*, 59–84.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171-191.

Grabe, W. (2001). Notes towards a theory of second language writing. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 39–58). Mahwah, NJ: Erlbaum.

Jones, R. L. (1985). Second language performance testing: an overview. In P. C. Hauptman., R. LeBlanc & M. B. Wesche (Eds.), *Second language performance testing* (pp.15-24). Ottawa: University of Ottawa Press.

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

Kellogg, R. T. (1988). Attentional overload and writing performance: Effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 355–365.

Kellogg, R. T. (1990). Effectiveness of prewriting strategies as a function of task demands.

American Journal of Psychology, 103, 327–324.

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

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18,* 219–240.

Linacre, J. M. (2005). *A user's guide to FACETS: Rasch-model computer programs [Software manual]*. Chicago, IL: Winsteps.com.

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.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

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

Mohan, B. A., & Lo, W. A. Y. (1985). Academic writing and Chinese students: Transfer and developmental factors. *TESOL Quarterly, 19*(3), 515–534.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Rafik-Galea, S. (2003). Enhancing writing ability through portfolios. *English Language & Literature Teaching, 9*(2), 1-15.

Schaefer, E. (2008). Rater bias patterns in an 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.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly, 27*(4), 657–675.

Song, M. S. (2002). The way to improve the English writing ability based on the performance assessment. *English Language & Literature Teaching, 8*(1), 165-197.

Uzawa, K. (1996). Second language learners' processes of L1 writing, L2 writing, and translation from L1 into L2. *Journal of Second Language Writing, 5*(3), 271–294.

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

Weigle, S. C. (2000). Investigating rater/prompt interactions in writing assessment: quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

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

APPENDIX

Analytic Scoring Rubric[1]

Level / Criteria	5 Advanced-High	4 Advanced –Low	3 Intermediate-High	2 Intermediate-Low	1 Novice
Content Logical development of ideas Main ideas, supporting ideas, and examples	Effectively addresses the topic and task, using clearly appropriate explanations, examples, and details	Addresses the topic and task well with using appropriate explanations, examples, and details	Addresses the topic and task using somewhat developed explanations, examples and details	Limited development in response to the topic and task using inappropriate explanations, examples and details	Questionable responsiveness to the topic and task with using no detail or irrelevant explanations
Organization The sequence of introduction, body, and conclusion Use of cohesive devices	Well organized and cohesive devices effectively used	Fairly well organized and cohesive devices adequately used	Loosely organized and incomplete sequencing; cohesive devices may be absent or misused	Ideas are disconnected and lack of logical sequencing inadequate order of ideas	No organization and no use of cohesive devices
Language in use Choice of vocabulary Register	Appropriate choice of words and use of idioms	Relatively appropriate choice of words and use of idioms	Adequate choice of words but some misuse of vocabulary or idioms	Limited range of vocabulary, confused use of words and idioms	Very limited vocabulary, very poor knowledge of idioms
Grammar Sentence-level structure	No errors, full control of syntactic variety	Almost no errors, good control of syntactic variety	Some errors, fair control of syntactic variety	Many errors, poor control of syntactic variety	Severe and persistent errors, no control of syntactic variety
Mechanics Punctuation/ Spelling Capitalization/Indentation	Mastery of spelling and punctuation	Few errors in spelling and punctuation	Fair number of spelling and punctuation errors	Frequent errors in spelling and punctuation	No control over spelling and punctuation

Examples in: English
Applicable Language: English
Applicable Levels: Secondary

[1] A modified version of Cohen's (1994) and Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey''s (1981) scoring criteria

Yousun Shin
Division of English Language and Literature
Pukyong National University
599-1 Daeyeon 3-dong, Nam-gu
Pusan, Korea 608-737
C.P: 010-4760-0914
Email: yousun-shin@pknu.ac.kr

Received in January, 2010
Reviewed in February, 2010
Revised version received in March, 2010