English Language & Literature Teaching, Vol. 13, No. 2 Summer 2007

Investigating Learners' Perception on Their Engagement in Rating Procedures

Ho Lee
(Chung-Ang University)

Lee, Ho. (2007). Investigating learners' perception on their engagement in rating procedures. *English Language Literature & Teaching, 13*(2), 91-108.

This study investigates learners' perception on their engagement in rating activities in the EFL essay-writing context. The current study aims to address the answers to the following research questions: 1) What attitude do students show about their participation in the rating tasks? and 2) which of three aspects (e.g. the degree of rating experience, the exposure to English composition instruction and learning, and proficiency level) significantly influences learners' rating activities? 104 EFL learners participated in the rater training session. After participants finished rater training session, they rated three sample essays and peer essays using the given scoring guide. Based on the analysis of survey responses that students made, students showed positive attitude toward their engagement in rating tasks. For research question 2, only L2 writing proficiency seriously affected students' perception on the rating tasks. Advanced level of subjects did not feel stressed by a grade of peers as low level of subjects did. They were also critical about the benefits of self- and peer-assessment, suggesting that a peer's feedback on their own essay was not so useful and that a self-rating does not fully help learners identify their writing proficiency.

[Self assessment/Peer assessment/Writing assessment]

I. INTRODUCTION

Despite the frequent use of self and peer assessment in language writing classroom, there have been relatively few attempts to understand learners as evaluators and/or raters. In EFL context, some researchers valued peer evaluation as one of interactive social activities in the process-oriented writing (Prochaska, 2005; Kim, 2006). These studies, however, do not sufficiently provide what aspects substantially influence learners' evaluation and what students think of during their evaluation.

In ESL context, while a great number of second language writing studies investigated peer-comment and peer-interaction in the process of revision, fewer articles have been concerned about peers' measurement activities during peer-rating. This is partly because extensive rating activities through self, peer, and expert introduce more variances than traditional assessment. For test-developers, who would reduce undesirable variables, students' participation in the rating procedure may threaten test reliability and test security. Many professionals presumed without question that a rating task is reserved for the professionals or experts in any context.

However, as some researchers suggested (Lee, 2005b; Luoma & Tarnanen, 2003), learners' engagement in rating tasks boosts learners to share the ownership of judgment with teachers. Further, Lee's study (2005a) suggested that learners would take responsibility for peers' learning in peer-rating context. Self- and peer-rating, if they are cautiously designed for the proper purpose, are strong meta-cognitive tools that allow learners to practice judgment skills.

These benefits of rating activities certainly may offer a great motif for second language professionals and researchers. If learners' participation in the rating activities deserves attention, further studies need to be conducted in succession. The current study was launched with the awareness of the needs, specifically focusing on the characteristics which might influence learners' judgment in self and peer rating contexts. The current study has the following research questions:

1) How do learners think of their participation in the rating tasks?
2) What aspects affect the learners' perception on their rating tasks?

II. REVIEW OF LITERATURE

The current study aims to investigate whether three factors-rater training, exposure to composition learning and teaching, and language proficiency-significantly affect the learner's perception on their engagement in rating activities. Unfortunately, it is rarely known about the relationship between learners' knowledge about English composition and the accuracy/faithfulness of rating tasks. On the other hand, the factors influencing expert raters' judgment have been extensively explored among second language writing researchers. As a starting point, this section starts with introducing the issues about rater training and its effect on rating performance.

A lot of second language writing studies stressed upon positive sides of a rater training program. Weigle (1994) found that even though rater training did not

completely wipe out raters' subjectivity, it softened the inter-rater differences. Shohamy, Gordon, and Kraemer (1992) asserted that rater training more significantly contributes to rater reliability than rater experience. Lumley and McNamara (1995) observed that rater training allowed novice raters to have self-confidence on their ratings. Weigle (1998), however, was cautious about the use of rater training and warned that only a single doze of rater training did not guarantee the long-term effects on rater reliability. All these studies assumed that raters were candidates for experts, ESL teachers, or professionals.

Stanley (1992) and Berg (1999) observed the beneficial effect of training on the depth and quality of feedback learners offered to their peers. Stanley (1992) guided students to go through coaching procedure in which they carried out role-playing, analyzed evaluation session, and learned the composition rule. The participants receiving training more vividly participated in the evaluation tasks and showed more productive comments on peers' essays. Berg (1999) also supported that while students receiving training reserved their feedbacks on local grammatical errors, they provided in-depth revision on organization and content in peer-evaluation. Both studies, however, did not configure the student engagement in the rating activities. The studies focused on their evaluative feedback to peer essays in a classroom situation than quantitative ratings that those who are qualified as experts did in a test context.

Since the studies have been rarely conducted about the students' participation in the rating procedure, there are demands for research which attempts to understand learners as raters in second and foreign language writing.

There are only a few studies about the relationship on psychological and personal characteristics of the raters in self- and peer-rating. According to AlFallay (2004), learners with low self-esteem and integrative orientation are the most accurate raters in self-assessment of oral presentation tasks. Learners with high self-esteem and with instrumental orientation were the least accurate raters. The agreement value between teacher-rating and peer-rating was higher for students with high anxiety and with integrative orientation than for those with high self-esteem and with low classroom anxiety. AlFallay suggested that high self-esteem yields the over-rating pattern especially in peer-assessment and, therefore, that learners with the positive traits (e.g. high self-esteem, low classroom anxiety) are not more reliable raters than those with negative side of the same trait.

Lee (2005a) suggested that students who marked themselves as high proficient ones showed greater rating confidence than less proficient classmates. According to his study, self-marked high Korean students were likely to recognize themselves as

accurate raters. He concluded that differing L2 proficiency is partly, but not fully associated with rating performance. In his further study (2005b), when learners rated peers' essays, they tended to concentrate on organization and content because of their lack of linguistic proficiency. The series of empirical studies evidenced that language proficiency is a crucial factor to affect the way students assess essays and their perception on rating tasks.

Up to now, the existing studies have partly illuminated the learners as a prospective raters in the second/foreign language learning contexts. These studies have also been rarely conducted about learner's perception about their participation in the rating tasks.

III. METHODS

1. Subjects and Setting

Initially, 104 Korean university students residing in Korea participated in the current study. Ninety one of 104 participants majored in English language education and/or literature, while only a few students specialized in other fields. Eighty three of the subjects are freshmen and sophomores, while only 21 subjects are juniors and seniors. Since a few students missed to mark their answers on some survey items, 99 to 102 responses per survey item were collected.

In particular, the current study introduced three categorical independent variables: the degree of rating experience, the exposure to English composition instruction and learning, and proficiency level. The first two were obtained by each student's responses to the questionnaire item 1 to 2 on the student background. Proficiency level was determined by the average holistic score of two expert raters on each student' essay.

In Table 1, these three variables were connoted as, respectively, 'Experience', 'Exposure', and 'Level'. The students who participated in rater training in the previous research were categorized as 'EER (Experienced in English Rating)'. 46 EER students were invited for the current study in order to examine long-term effect of rater training. `NER (Not-experienced in English Rating)' did not take part in the rater training before. Participants under the category of 'EEC (Exposure to English Composition)' were those who learned how to write an English essay in a minimum of 10 hours of classroom time, while the rest of the students were called 'NEC (Non-exposure to English Composition)'.

TABLE 1
Frequency of Each Subgroup

	Experience		Exposure		Level		
Statistics	EEC	NEC	EER	NER	Low	Mid	High
f	46	58	45	59	44	53	7
%	44	56	43	57	42	51	7

Regarding the proficiency level, 'Low' indicates less than '2.5', 'Mid', from '2.5' to '3.5', and 'High', more than `3.5'. From Table 1, the distribution of the students was found to be positively skewed. Even though the majority of students ranked in the intermediate level, 44 out of 104 students received less than score 2.5. This was probably because only a few students have experienced to write an argumentative English essay for the test purpose.

2. Instrumentation

Several instruments were used for the current research. Of these, the questionnaire development and benchmark design deeply influenced the result of the current study. This section just included brief explanation about the development and use of the two instruments.

In making a scoring benchmark, the researcher adapted several existing rubrics such as Willard-Traub, Decker , Reed, and Johnston (1999), TOEFL writing scoring guide (ETS, 2000), and an ESL placement test feature analysis form used in the University of Illinois at Urbana-Champaign. The present benchmark suggested the scoring guidelines for four analytic features (organization, content, language use, plagiarism) and for one holistic judgment. Each of analytic features contained more than three semi-features. For example, students were induced to consider three aspects (intro, body, and conclusion) in rating organization. Each semi-feature provided a descriptor that indicated the level of quality of the text. An excerpted benchmark is provided in Appendix 1.

Survey items (Appendix 2) were developed with a categorical scheme. The survey was divided into three parts: students' background, rating process, and student's perception. Considering the object of this research, the researcher only included background items (from item 1 to 2) and perception items (from item 30 to 54) in this paper. Therefore, total 27 survey items were analyzed for this study.

Survey items in the section of background were YES/NO type of questions. Of four YES/NO questions, item 1 and item 2 were developed to identify subgroups. Twenty-

five items in the part of perception had 5 Likert-scale from strongly disagree to strongly agree, whereas six questions were developed in the type of multiple choices. In order to enhance the quality of the questionnaire, five experts and two pre-selected undergraduate students examined the survey items before the researcher conducted the main experiment.

3. Procedure

On the day of data collection, undergraduate EFL students underwent the rater training session. At first, a researcher suggested clear guidelines about the scoring criteria, showing the model about how an expert raters rate essays. Then, subjects rated three sample essays using the scoring benchmark. After they finished their rating activities, they exchanged their scores and the rating procedure with their partners pre-assigned by the researcher. Those who finished negotiation with their partners were led to assess their own essays and, respectively, their partners' essays. Then, they provided their rating scores and oral/written feedback to each other, As soon as they finished rater training session and multiple rating activities, they were requested to fill out a survey.

Two expert raters were invited to rate the subjects' essays. The average scores decided by two experts were used to discriminate advanced level from low and mid level.

4. Data Analysis

In order to view the overall pattern of students' responses to Likert-scale items, a factor analysis is conducted. The factor analysis also aims to provide an analysis scheme with which students' perception on rating tasks is systematically reported. Table 2 shows that, based on the varimax rotation, six factors are initially identified from nineteen items. Factor 6 is excluded from the subsequent analyses since it contains only one item, item 39, although the Eigenvalue value is over 1.00. Alternatively, item 39 is placed into factor 5 since the item is thematically similar to the items of factor 5. That is also true for item 47. The survey item 47, which is initially included in factor 3, moves to factor 1 for the following analyses. In addition, item 32 is excluded from factor 4 because at a first glace the item measures a different construct. The current study does not include the analysis of item 32.

In creating these divisions, five factors are formed. Each of the factors is named as the following: 'overall evaluation' for factor 1, 'sensitivity to external rating' for factor 2, 'a rating task as a means of self-diagnosis' for factor 3, 'perception on group rating' for factor 4, and 'benefits of rating tasks' for factor 5. Factors1, 3, and 5 include four items each (item 45, 46, 47, and 48 for factor 1; item 37, 38, 41, and 42 for factor 3; item 36, 39, 43, and 44 for factor 5), while each of the other factors contains three items (item 30, 31, and 40 for factor 2; item 33, 34, and 35 for factor 4). Each of these factors functions as a title of each forthcoming section.

Finally, two remaining entries ('comparisons of self- and peer-ratings on accuracy' and 'expected proficiency level of a peer') are identified from the analyses of the rank-order type of items. Therefore, the survey analyses consist of seven sections.

TABLE 2

Factor Loading of Survey Items

	Factors				
Survey item description	1	2	3	4	5
45. Overall, rating tasks were easy.	**.768**	-.134	-.077	.138	.035
46. Overall, I accurately rated essays.	**.773**	-.015	.054	-.180	.086
48. To serve as a complementary rater.	**.646**	.476	-.069	.197	.000
[a]47. Overall, rating tasks were interesting.	.499	-.083	.375	**.514**	-.171
31. To be suspicious of experts' grades.	.096	**.845**	.058	-.256	.063
30. I am interested in experts' assessment.	.067	**.779**	-.243	.010	.051
40. I felt shamed by the partner's grade.	-.072	**.648**	.102	-.028	-.174
37. To know my proficiency objectively.	.037	-.080	**.875**	-.069	.066
41. A peer's feedback was useful.	-.203	.229	**.578**	.105	-.096
38. To compare the peer's essay to mine.	.242	.056	**.469**	-.039	.279
42. To detect my strengths/weaknesses.	-.267	.345	**.392**	.130	.142
33. It is easy to understand the benchmark.	.098	-.057	-.148	**.878**	-.060
35. I was satisfied with group rating.	-.044	-.132	.201	**.717**	.048
34. I fully discussed rating with a partner.	.010	-.009	-.425	**.443**	.120
[b]32. L2 proficiency is crucial for the rating.	-.170	.177	-.056	**.407**	-.076
43. To be critical of the expert's rating.	.069	.044	-.009	-.219	**.925**
44. Ratings help me improve writing.	.062	-.110	.151	.183	**.790**
36. Rater training helped me learn rating.	-.213	-.068	-.012	.288	**.466**
[c]39. I took responsibility in the peer-rating.	.041	-.030	.072	-.134	-.299
Eigenvalue	3.77	2.26	1.69	1.50	1.30
% of variance	19.82	11.91	8.89	7.87	6.81
Cumulative %	19.82	31.73	40.62	48.49	55.30

Note. Bold = the highest loading for each variable.

[a] Item 47 moved from factor 4 to factor 1. [b] Item 32 was excluded from factor 4 and was separately investigated in 'expected proficiency level of a peer'. [c] Item 39 was included in factor 5 as factor 6 was excluded.

From Table 3 to Table 7, the response categories such as 'positive', 'neutral', and 'negative' correspond to each of the identified subgroups. A 'positive' group consists

of students who marked '4' and '5' (from agree to strongly agree) in the Likert-scale questionnaire items, while those in a 'negative' group were respondents marking '1' and '2' (from strongly disagree to disagree). Participants checking '3' in the Likert-scale items were classified as a 'neutral' group.

IV. RESULTS AND DISCUSSION

1. Overall Evaluation

As indicated in Table 3, there are no significant group differences in the aspects of overall evaluation. Every subgroup tends to evenly respond that a rating task is not easy. What is more, 73% of the total students are not confident in the accuracy of their ratings. A small majority of the students express their interest in the rating task and their acceptance of the offer of serving as a complimentary rater. On the basis of the findings, while a rating is viewed as a cognitively challenging task, it is a motivating form of learning.

TABLE 3

Responses to Survey Items About Overall Evaluation

			Exposure		Experience		Level		
Item	Response	*f*	EEC	NEC	EER	NER	Low	Mid	High
45	Positive	27	11	16	13	14	9	14	4
	Neutral	33	16	17	16	17	11	20	2
N=101	Negative	41	16	25	17	24	21	19	1
46	Positive	23	14	9	10	13	7	13	3
	Neutral	51	22	29	19	32	19	29	3
N=101	Negative	27	10	17	14	13	15	11	1
47	Positive	43	21	22	20	23	15	24	4
	Neutral	34	16	18	12	22	17	14	3
N=101	Negative	24	9	15	11	13	9	15	-
48	Positive	41	22	19	21	20	12	24	5
	Neutral	28	12	16	9	19	13	14	1
N=101	Negative	32	12	20	13	19	16	15	1

2. Sensitivity to External Rating

The result from the table 4 reveals that more than half of students are sensitive to external ratings. Sixty seven of 101 subjects are concerned about the way their academic performance is graded, regardless of the subgroups they are involved in. On

the other hand, exposure to composition differentially influences student's responses to the item 31, "I was often suspicious of a teacher's or a professor's grading criteria". EEC group is likely to be more critical of a teacher's grading, χ^2 (2, N = 101) = 10.18, p< .01. On item 40, high proficient subjects are less vulnerable to the emotional effect triggered by a peer-rating than those who are less proficient, χ^2 (4, N = 100) = 9.32, p< .05.

TABLE 4

Responses to Survey Items About Sensitivity to External Rating

			Exposure		Experience		Level		
Item	Response	*f*	EEC	NEC	EER	NER	Low	Mid	High
30	Positive	67	34	33	25	42	23	39	5
	Neutral	20	9	11	10	10	9	10	1
N=101	Negative	14	2	12	7	7	9	4	1
31	Positive	57	33	24	24	33	23	32	2
	Neutral	32	10	22	13	19	12	16	4
N=101	Negative	12	2	10	5	7	6	5	1
40	Positive	54	23	31	22	32	23	30	1
	Neutral	28	14	14	12	16	13	14	1
N=100	Negative	18	8	10	7	11	5	8	5

Note. χ^2 (4, N = 100) = 9.32, p< .05 for Level in item 40. χ^2 (2, N = 101) = 10.18, p< .01 for Exposure in item 31.

3. Rating Tasks as a Means of Self-Diagnosis

Table 5 shows that students see rating activities as having a positive effect on their writing studies. The analysis of students' responses to item 37 supports that self-assessment fosters independent learning (Oscarsson, 1989). Fifty-eight of the subjects claim self-rating increases their awareness of their current proficiency level. They regard a rating task as a useful means of improving their writing skill. However, advanced level of students is skeptical against the usefulness of peer-rating tasks. They also give the least positive responses on the item 37, "Self-ratings helped me objectively identify my overall writing proficiency."

TABLE 5

Responses to Survey Items About Self-Diagnosis

			Exposure		Experience		Level		
Item	Response	*f*	EEC	NEC	EER	NER	Low	Mid	High
37	Positive	58	24	34	22	36	21	36	1

	Neutral	36	18	18	17	19	17	14	5
N=99	Negative	5	2	3	2	3	3	1	1
38	Positive	62	30	32	28	34	25	33	4
	Neutral	28	12	16	9	19	14	12	2
N=99	Negative	9	2	7	4	5	3	5	1
41	Positive	49	22	27	24	25	22	25	1
	Neutral	36	17	19	10	26	13	21	2
N=99	Negative	14	6	8	7	7	5	5	4
42	Positive	53	23	30	23	30	26	26	1
	Neutral	39	21	18	17	22	11	23	5
N=101	Negative	9	2	7	3	6	4	4	1

Note. $\chi^2(4, N = 99) = 12.10$, $p < .05$ for Level in item 41. $\chi^2(4, N = 99) = 18.55$, $p < .01$ for Level in item 37.

4. Group Rating During Rater Training

Table 6 descriptively illustrates analyses of students' perception on group rating per subgroup. There are no significant differences among the subgroups for item 30, 34, 35. At the beginning of the rater training session, students seem to not fully understand the scoring rubric. As some students indicate in their comments, most students are overwhelmed by a number of score descriptors. Hence, they are under time pressure since they are requested to read the benchmarks line-by-line in a given time. The time pressure also affects the negotiation process so that students do not have sufficient talk with their partner in score decision for their lack of knowledge of the benchmark. On the other hand, fifty students express their satisfaction with the score decision made by negotiation.

TABLE 6

Responses to Survey Items About Group Rating

			Exposure		Experience		Level		
Item	Response	f	EEC	NEC	EER	NER	Low	Mid	High
30	Positive	37	19	18	18	19	14	19	4
	Neutral	48	21	27	17	31	21	25	2
N=100	Negative	15	5	10	7	8	6	8	1
34	Positive	31	15	16	16	15	10	17	4
	Neutral	53	25	28	20	33	23	27	3
N=102	Negative	18	6	12	7	11	9	9	-
35	Positive	52	24	28	26	26	19	28	5
	Neutral	44	21	23	15	29	20	23	1
N=102	Negative	6	1	5	2	4	3	2	1

5. Benefits of Rating Tasks

As indicated in Table 7, participants make positive responses regarding the use of the rating task. Sixty two of the students learn how to rate an English essay through a rater training session. Fifty-four students also expect that the rating task will help them improve their writing skill. On the basis of the students' responses on item 43, a rating task enables fifty-six students to acknowledge social surroundings more keenly than before. The advanced level of students, however, reserves their positive response. A large number of students (68 out of 100) have a heightened sense of responsibility when they conduct a peer-rating. These positive responses on item 39 show how beneficially a collaborative rating is brought into action. Sharing the responsibility for assessment with each other, students generate a more discerning attitude toward their peers through a peer-rating.

TABLE 7
Responses to Survey Items About Benefits of Rating Tasks

			Exposure			Experience		Level		
Item	Response	*f*	EEC	NEC		EER	NER	Low	Mid	High
36	Positive	62	28	34		25	37	23	35	4
	Neutral	36	16	20		17	31	16	17	3
N=101	Negative	3	1	2		1	2	3	-	-
39	Positive	68	33	35		25	43	24	40	4
	Neutral	26	12	14		13	13	14	9	3
N=100	Negative	6	-	6		3	3	4	2	-
43	Positive	56	25	31	31	23	33	23	31	2
	Neutral	33	16	17		14	19	10	19	4
N=101	Negative	12	5	7		6	6	9	2	1
44	Positive	54	23	31		20	34	24	27	3
	Neutral	31	17	14		14	17	12	17	2
N=101	Negative	16	6	10		9	7	6	8	2

6. Comparison of Self- and Peer-Rating on Accuracy

Table 8 compares self- and peer-ratings on accuracy per subgroup. All subgroups show similar pattern for this aspect. Approximately 46% of survey respondents think that a peer-rating is more accurate than a self-rating, while the rest of the students accept that their own essay scores are at least as accurate as peer-ratings.

TABLE 8
Responses to Survey Items About the Comparisons of Self- and Peer-Rating

			Exposure		Experience		Level		
Item	Response	*f*	EEC	NEC	EER	NER	Low	Mid	High
51[a]	Yes	19	6	13	8	11	11	6	2
	Same	34	19	15	14	20	14	19	1
N=99	No	46	20	26	19	27	16	26	4

[a] Item 51: "A self-rating was more accurate than a peer-rating".

7. Expected Proficiency Level of a Peer

According to Table 9, 64 of 100 students prefer a more proficient person than themselves as a peer-rater. On the other hand, only 17 of the 64 respondents want to rate the essay of a more proficient peer. This indicates that students are concerned about their less proficient levels, which may cause negative impact on the rating accuracy. Overall, students do not show a dominant preference to a specific level of a peer's essay.

TABLE 9
Cross-Tabulation of Students' Responses on the Favored Proficiency Level of Peers

	Item 53[a]				
Item 54[b]	Above self	Same as self	Below self	I don't care	Total
Above self	17	4	1	-	22
Same as self	19	18	2	-	39
Below self	18	2	2	-	22
I don't care	10	-	-	7	17
Total	64	24	5	7	100

Note. Item 53[a]: " If an anonymous peer rates your essay, which proficiency level of peer do you prefer in comparison with your current level?". Item 54[b]: "If you rate an anonymous peer's essay, which proficiency level of peer do you prefer in comparison with your current level?".

Table 10 includes students' comments about the expected proficiency level of a peer-rater and a peer essay. The majority of students' comments demonstrate a strong preference toward a more proficient peer-rater. The participants' opinions are imbued with the expectation that a highly proficient person better conducts a rating task. When students are requested to rate a peer's essay, they do not adhere to the preference for an advanced level of the peer. 'Above self' and 'I don't care' respondents clearly acknowledge that a rating task is not simply a measurement of someone, but is also a learning activity. That is, they attempt to learn something from others' essays.

On the other hand, the respondents of 'Below self' and 'Same as self' are highly concerned with rating accuracy.

TABLE 10
Excerpts From Students' Comments on the Preferred Level of a Peer-Rater and Peer-Essay

Item No.	Respondent's Choice	Selected comments
53	Above self	"A higher proficient peer better points out my mistakes and errors." "Because a peer accurately rates my essay in grammar and vocabulary" "A high-level peer's rating is reliable." "A proficient peer seems to be more severe and objective." "I will learn something from the proficient peers." "A proficient peer shows the great command of word choice." "It is possible for the peer-rater to give a productive comment."
	Same as self	"I will not feel shame."
	Below self	"I look like a better student than a peer."
	I don't care	"Only if the rubric is given, I will not care."
54	Above self	"I will learn something by rating the essay of a higher proficient peer." "I want to learn polished and refined English sentences." "It is easy to detect my writing errors by reading a peer's essay."
	Same as self	"It is difficult to understand an advanced peer's essay." "Only after I compare my essay to a peer's essay, I am able to give substantial feedback."
	Below self	"It will be easy to rate an essay." "I am able to give comprehensive and correct feedback." "An accurate rating is possible only when a peer's level is lower than mine."
	I don't care	"It will be very helpful to read a variety of essays. One should read many essays, regardless of the varying qualities."

V. LIMITATIONS, SUGGESTIONS AND CONCLUSION

The survey analysis suggests that students showed a variety of perspectives on the rating tasks. A majority of students saw the positive side of the rating tasks, believing that the rating tasks would enhance the critical awareness of the surrounding social context, independent learning, and the sense of responsibility for a peer classmate's learning.

This study also found that both rating experience and exposure to the composition teaching and learning did not seriously affect students' perception on the rating tasks. On the other hand, students made differential reaction to some survey items dependent upon their varying proficiency levels. Advanced level of subjects did not feel stressed

by a grade of peers as low level of subjects do. They are also critical about the benefits of self- and peer-assessment, suggesting that a peer's feedback on their own essay was not so useful and that a self-rating does not fully help learners identify their writing proficiency. Intermediate and low proficient students were less skeptical about the effect of peer- and self-rating.

Students reported that they preferred more proficient peers than themselves as a peer-evaluator. They were convinced that advanced learners would give more accurate and truthful feedback. That is, learners thought of high language proficiency as the required, and at least primary condition for the right judgment. Yet, they displayed varying responses about the preferred quality of peer essays they should rate. Learning-focused students are concerned with how good peer writers write an essay, while accuracy-concerned students apprehended that their low proficiency would fail to lead to reliable judgment.

The present study was born with some limitations in the research procedure. First, a single administration of the survey per each student may provide partial information about what students really think. Some students make differential responses dependent upon the content, their emotional mood, the question type, and etc. In a further study, multiple surveys conducted for prolonged period (i.e the whole semester) are requested to acquire more reliable data.

Second, it should be noted that students' comments may possibly distorted in the translation process. I allowed the learners to use their mother tongue, Korean, in order to reduce learners' stress. Therefore, the researcher's subjective judgment may bring into action in the translation process from Korean to English.

Third, the small size of advanced level of subjects may undermine the dependability of the results. Only seven learners' responses were not sufficient to make a general statement about behavioral patterns for advanced learners.

Most of all, the result is not buttressed by further qualitative analysis. Because the current study dominantly used a quantitative analysis to identify the overall pattern from the students' responses, in-depth inquiry was nearly impossible. Only comments a few students took down in the survey were not enough to read respondents' ideas. A further study needs to employ the series of interviews to several students after they complete rater tasks.

Until now, the current study evidenced that students' participation in the rating process invites interesting issues from the perspectives of learners. The partial-credited and sometimes uncertain findings of this study urgently need to be confirmed by the series of further studies.

REFERENCES

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System, 32*(3), 407-425.

Berg, E. C. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing, 8*(3), 215-241.

ETS. (2000). *Test of English as a foreign language (TOEFL)*. Princeton, NJ: Educational Testing Service.

Kim, J. T. (2006). Filling the understanding gap of the misplacement of ESL learner's writing placement test. *English Language & Literature Teaching, 12*(3), 147-166.

Lee, H. (2005a). Examining the effect of L2 proficiency on a peer-rating procedure in the EFL context. *Foreign Languages Education, 12*(4), 211-234.

Lee, H. (2005b). Investigating the student's score decision in the EFL essay-writing context. *English Language Teaching, 17*(3), 155-179.

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

Luoma, S., & Tarnanen, M. (2003). Creating a self-rating instrument for second language writing: From idea to implementation. *Language Testing, 20*(4), 440-465.

Oscarsson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6*(1), 1-13.

Prochaska, E. (2005). Korean EFL university students' evaluation of peer review interactions: A social model for evaluating the writing process. *English Language & Literature Teaching, 11*(2), 51-66.

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

Stanley, J. (1992). Coaching student writers to become effective peer evaluators. *Journal of Second Language Writing, 1*(3), 217-233.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

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

Willard-Traub, M., Decker, E., Reed, R., & Johnston, J. (1999). The development of large-scale portfolio placement assessment at the University of Michigan: 1992-1998. *Assessing Writing, 6*(1), 41-84.

Appendix 1
Benchmark

Directions: The following describes each scoring descriptors from score 1 to score 5 across 4 features. Please read carefully each scoring descriptors and give one of 5 step-scaled scores (1, 2, 3, 4, 5) in terms of each of four features.

Score 1

Feature	Analytic Feature	Descriptors
Holistic	Organization Content Language Use Plagiarism	No organization of ideas; Insufficient length to ascertain organization; No transitions; Little or no detail, or irrelevant specifics; A vague, general telling of the idea; Sentence variety and complexity not present; Extremely bad grammar; Totally incomprehensible; Serious mechanical/punctuation problem; Serious problems with focus; Many times off topic.
Org	Intro, Body, Conclusion	No organization of ideas; No intro/body/conclusion; More than two of three structures (intro/body/conclusion) are not present.
	Cohesion within paragraphs	No transition between sentences; No thesis statement.
	Coherence b/n paragraphs	No continuity (flow of idea) between paragraphs.
	Repetition	Repetitions of many words/idioms within paragraphs and/or within sentences; Some sentences with the same meaning and almost the same form are recurrent; Likely to be repetitive in rhetoric level.
Content	Relevance to Topic	Totally, off-topic.
	Support	No support or elaboration of ideas; Few detail.
	Writer's idea	A vague, general telling of the fact without student's own idea.
	Repetition	Repetitions of many words/idioms within paragraphs and/or within sentences; Some sentences with the same meaning and almost the same form are recurrent; Likely to be repetitive in rhetoric level.
	Plagiarism	Majority copied without citation from source information.

Languag e Use	Grammar Accuracy	Extremely bad grammar; Totally incomprehensible; Grammar errors are present in every aspect; Serious mechanical/punctuation problem; Many grammatical errors hamper the meaning.
	Spelling	Many spelling errors.
	Appropriate Word Choice	Some critical lexical errors each paragraph; Critical misuse of words/idioms causes incomprehensibility.
	Syntactic Structure	Sentence variety and complexity not present; Incomplete sentences.

Appendix 2

Questionnaire

Name:_____ Major:_____ Email:_____ Academic Year: _____

Directions: Please read each statement carefully and choose one of the choices that best represents your opinion. Please be honest in filling in the questionnaire.

Background

01. I have been exposed to English composition instruction and learning more than 10 hours for this semester. Yes No
02. I participated in the pilot study in June 2004. Yes No

Perception on Rating Tasks (Pre- and Post- Task Level)

Likert-scale Questions

30. Usually, I am interested in how my academic performance is graded. 1 2 3 4 5
31. I was often suspicious of a teacher's or a professor's grading criteria. 1 2 3 4 5
32. English proficiency is the most significant determinant for rating accuracy. 1 2 3 4 5
33. It was easy to understand the scoring guide. 1 2 3 4 5
34. During rater training, I have sufficient negotiation with my partner in making decision. 1 2 3 4 5
35. I was satisfied with the score decision that I and my partner made. 1 2 3 4 5
36. Rater training helped me learn how to assess the essay accurately. 1 2 3 4 5
37. Self-ratings helped me objectively identify my overall proficiency of writing. 1 2 3 4 5
38. During peer rating, I compared the quality of the peer's essay to that of my essay. 1 2 3 4 5
39. I felt the sense of responsibility in rating a peer's essay. 1 2 3 4 5
40. I felt ashamed by the scores a partner gave on my essay. 1 2 3 4 5
41. A peer's feedback on my essay was useful. 1 2 3 4 5
42. Ratings helped me detect my strengths and weaknesses in writing. 1 2 3 4 5
43. I became to have a critical view about the ways an expert rates an English essay. 1 2 3 4 5
44. Rating tasks will help me improve English writing proficiency. 1 2 3 4 5
45. Overall, rating tasks were easy. 1 2 3 4 5
46. Overall, I accurately rated essays. 1 2 3 4 5
47. Overall, rating tasks were interesting. 1 2 3 4 5
48. I will voluntarily serve as a complementary rather in a large-scale essay test, if I am offered. 1 2 3 4 5

Comparison Questions

49. During the rater training session my ratings on the sample essays were more accurate than experts' ratings. Yes Same No
50. During rater training, a group rating is more reliable than an expert's rating. Yes Same No
51. A self-rating was more accurate than a peer-rating. Yes Same No

52. A peer-rating was more difficult than a self-rating. Yes Same No

Multiple Choice & Short Answer

53. If an anonymous peer rates your essay, which proficiency level of peer do you prefer in comparison with your current level?
    a. More proficient peer b. The same proficient peer
    c. Less proficient peer d. I don't care
    Why?
54. If you rate an anonymous peer's essay, which proficiency level of peer do you prefer in comparison with your current level?
    a. More proficient peer b. The same proficient peer
    c. Less proficient peer d. I don't care
    Why?

Examples in: English

Applicable Language: English

Applicable Levels: College

Ho Lee
Dept. of English Education
Chung-Ang University
221, Heuk-Seok Dong, Dong-Jak Gu
Seoul, 156-756, South Korea
Tel: (02)820-5393 / C.P.: 010-2354-4479
Email: holee@cau.ac.kr

Received in April, 2007
Reviewed in May, 2007
Revised version received in June, 2007