English Teaching, Vol. 60, No. 3, Fall 2005

Validity of Self- and Peer-ratings in an EFL Essay-writing Test

Ho Lee
(University of Illinois at Urbana-Champaign)

Lee, Ho. (2005). Validity of self- and peer-ratings in an EFL essay-writing test. *English Teaching, 60*(3), 195-219.

The current study investigates the validity of a self- and peer-rating in a second language (L2) essay-writing test in Korea. The current study suggests that the Korean EFL students whose major is English education should conduct the rating tasks that they will encounter in the post-secondary school in their future career. Each of 104 Korean EFL learners at the intermediate level sequentially rated three sample essays, his/her own essay, and his/her partner's essay after he/she was provided with the same benchmark as two experts would refer to. Each essay was rated on a 5-scale in terms of 4 writing features (organization, content, language use, and holistic feature). After students finished the ratings tasks, the students' ratings were compared to expert ratings, using Multi-facet Rasch model. As a result, peer raters were so lenient that they showed unpredictable rating pattern. This research is dedicated to seek an innovative testing system in which trained students participate in a test as a complementary rater specifically in the EFL context.

I. INTRODUCTION

A learner has been typically recognized as a test-taker when simply reacting to an item on a test. Even though a learner is claimed to construct his/her own knowledge, second language (L2) testers tend to underestimate learners' active involvement in a test. Furthermore, the paradigm that the quality of a test is determined primarily by psychometric characteristics of the testing group leads language testers to disvalue feedback from test takers in English language testing practice and testing validation processes (Hamp-Lyons & Lynch, 1998; Shin, 2003).

Some language testing professionals, while acknowledging limitations of a tester-oriented testing system, have suggested the administration of a collaborative assessment procedure such as self- and peer-assessment (Luoma & Tarnanen, 2003; Ross, 1998;

Shohamy, 2001). The collaborative assessment procedure as an alternative form of assessment takes into account learners' characteristics, experiences, backgrounds, talents, needs, and language proficiency (Ekbatani, 2000). It could also embrace the process during which learners share the operational assessment criteria with expert raters. That is, this allows for greater focus to be placed on the role of the test-taker during the assessment procedure.

However, only a few studies have been conducted on collaborative assessment procedures in second language writing contexts, even though collaborative learning is an important issue in the general educational field. This seems to be partly because of the rooted belief that the ownership of assessment belongs to a qualified, native, and expert rater. It is also said that if the consequences of collaborative assessment become an integral part of assessment itself, collaborative assessment may yield negative effects on the psychometric sense of reliability due to its rich variance.

Such a psychometric approach, however, has been recently criticized for the following reasons. Psychometric reliability should be simply viewed as a constituent of validity since the current stream of language testing has posed validity as the most important principle. Along with the unitary concept of validity, the upheaval of philosophical trends, i.e. hermeneutics, post-modernism, and constructivism, leads to criticism of existing positivistic stances.

Under the anti-naturalistic perspectives, interests in the learner's role in the test have been raised. Collaborative rating should be a well-timed topic in that it invites test takers to raise their voices in test construction and implementation.

II. LITERATURE REVIEW

1. Self-Assessment

Self-assessment refers to the involvement of students in identifying criteria to apply to their works, and making judgments about the extent to which they have met these criteria and standards (Boud, 1991). Self-assessment increases students' responsibility not only for their learning but also for the assessment of their learning processes and products (Luoma & Tarnanen, 2003; Lynch, 2001; Shepard, 2000). In part, self-assessment is closely connected with the concept of autonomy associated with independence, self-fulfilment, and self-monitoring (Littlewood, 1996; Rivers, 2001) or with the locus of control (Pierce, Swain, & Hart, 1993).

Self-assessment is closely related with alternative paradigms of assessment (Ekbatani, 2000; Luoma & Tarnanen, 2003). First of all, the ownership of assessment is allowed for

both learners and teachers/guides. Second, the goal of the self-assessment is to help learners self-monitor the developmental sequences of their learning (Lynch, 2001). Consequently, such a process-oriented assessment combines learning and assessment instead of separating them. In addition, the self-assessment score reflects not only performance but also metacognition, ability to assess one's own cognition (Rivers, 2001). Likewise, self-assessment fosters evaluative attitudes in learners by means of raising learners' awareness of their goals and achievements (Luoma & Tarnanen, 2003; Oscarson, 1989).

A number of research papers on self-assessment have looked at self-ratings' comparability with expert ratings (Janssen-van Dieten, 1989; LeBlanc & Painchaud, 1985; Pierce et al., 1993; Strong-Krause, 2000). LeBlanc and Painchaud (1985) found positive correlations between a self-assessment instrument and a standardized English proficiency exam. Luoma and Tarnanen (2003) verified that the self-ratings for the second language writing were fairly comparable with teacher assessment in the context of self-marked tests. Blanche and Merino (1989) suggested that learners could accurately assess their own learning when the self-test items described in behavioral terms contained the concrete linguistic context that the learners experienced. Bachman and Palmer (1989) analyzed the responses of a self-rating questionnaire and found that self-ratings are reliable and valid measures of language ability through the use of multi-trait and multi-method (MTMM) and confirmatory factor analysis (CFA).

In contrast, some researchers have proposed that self-assessment did not reveal students' language proficiency exactly. Pierce et al. (1993) found that the sets of self-assessment of language proficiency did not significantly inform objective measures of language proficiency. In a larger study that included assessment for all language skills, Janssen-van Dieten (1989) found there was no significant correlation between the self-assessment instrument and the proficiency test.

Along with the arguments about comparability of self-rating, there is a controversy concerning overestimation versus underestimation. According to Bachman and Palmer (1989), students better recognized the area in which they have difficulty than they did in the relatively easy area. However, Davidson and Henning (1985) reported the opposite results that students tend to exaggerate their own ability. Another study reported that self-raters have a tendency to rate themselves fairly high on the scale, even if some had only been learning Finnish for a few months (Luoma & Tarnanen, 2003). Some researchers suggested that anxious students tended to underestimate their competence relative to less anxious students, who tended to overestimate their competence (MacIntyre, Noels, & Clement, 1997; Stefani, 1994).

There are also various views on the factors causing the accuracy/inaccuracy of self-assessment. Second Language (L2) writing researchers identified sources of variances

influencing self-assessment such as the wording on the questionnaire (Bachman & Palmer, 1989; Ross, 1998), the language skills being assessed (Strong-Krause, 2000; Ross, 1998), the level of proficiency of the students (Heilenman, 1990; Rivers, 2001), the cultural background of the students (Strong-Krause, 2000), the type of self-assessment task (LeBlanc & Painchaud, 1985; Pierce et al., 1993; Ross, 1998; Stong-Krause, 2000), and both linguistic skill and material (Blanche & Merino, 1989).

A few studies dealt with the level of proficiency as one of the main factors affecting the accuracy of self-rating. Heilenman (1990) found that self-rating was higher in less proficient students than in more proficient students. According to Rivers (2001), experienced language learners with high proficiency tend to be autonomous. In other words, they utilized cognitive or metacognitive strategies with better self-monitoring ability and, thereby, performed self-assessment successfully.

In terms of the language skill being assessed, Strong-Krause (2000) reported that students assessed their speaking and writing skills more accurately than they did their reading skills. In contrast, Ross (1998) found that learners have more difficulty in estimating their own speaking and writing skills than their own reading and listening skills. In addition, Ross expressed the concern that flexible interpretations of the ordinal scale caused the inaccuracy of self-ratings.

Some research papers examined the type of self-assessment tasks and its effect on accuracy. Ross (1998) suggested that testing items for self-assessment should assess the second language skills used in contexts that ESL students have actually experienced. Further, Ross mentioned that accuracy is affected by the degree of experience the learners have in the self-assessment context. He pointed out the fact that "when the criterion is one that does not invoke episodic memory, the self-assessor may have to rely on a recollection of his or her own general proficiency to make the assessment. Subject may resort to relativity or be influenced by self-flattery (p. 16)." According to Pierce et al. (1993), "the more specific and focused a self-assessment instrument is, the greater the likelihood that there will be higher correlations with objective measures of proficiency" (p. 39). Strong-Krause (2000) supported that the self-rating involved in the most specific tasks quite predicted the results of the placement exams. Other researchers pointed out the need for creating concrete linguistic situations where the learners can self-assess their communicative ability (LeBlanc & Painchaud, 1985; Luoma & Tarnanen, 2003; North, 2000).

In summary, many studies take the positive stance on the reliability and validity of self-assessment (Bachman & Palmer, 1989; Blanche & Merino, 1989; Oscarson, 1989). Self-assessment encourages learners to trigger meta-cognitive knowledge and high-level thinking such as problem-solving strategies. Learners are then autonomous as they control their own learning with their own responsibility. Psychometric validity and reliability of

self-assessment, however, cannot be easily gained. Test practitioners should control the relevant variances such as a learner's proficiency level, type of the self-assessment task, and so on. If learners with little rating experience are engaged with a vaguely described task, they will show low concurrent validity with important criterion variables. Ross pointed out the appropriate use of self-assessment and stated "depending on measurement needs and logistical constraints, self-assessment may be viewed as providing too cloudy a picture of proficiency for some testing decisions, e.g., candidate selection, or may be viewed as sufficiently accurate for other 'low stakes' decisions, e.g., placement within programs or rough-and-ready needs analysis instruments" (Ross, 1998, p. 12). Self-assessment, if used in the suitable conditions and contexts, is valuable in enhancing learner's own knowledge.

2. Peer-Assessment

Peer-assessment has gained a lot of interest as an innovative form of assessment. Topping and Smith (2000) considered peer-assessment as "an arrangement for peers to consider the level, value, worth, quality or successfulness of the products or outcomes of learning of others of similar status" (p. 150). Peer-assessment can be summative grading and informative reviews as a type of feedback. That is, peer-assessment encompasses quantitative measurement and qualitative verbal feedback. The first is referred to as peer-grading, or peer-rating. The latter is called peer comment, or simply peer feedback in this paper. Even though the peer comment has been widely discussed among L2 writing researchers, the peer-grading is rarely explored in this area. Therefore, this paper will mainly address issues of peer-grading discussed in general educational disciplines.

A few studies report a high correlation between peer and expert ratings (Hugh & Large, 1993; Stefani, 1994). Stefani (1994) concluded that learners provide very reasonable judgments to their peers in the context of the clearly defined and well-guided tasks. On the other hand, a majority of studies found that peer-grading is not sufficiently reliable to be used to supplant teacher grading (Cheng & Warren, 1999; Falchikov, 1995; Magin & Helmore, 2001; Mowl & Pain, 1995). Mowl and Pain (1995) examined how self-assessment, peer-assessment, and tutor assessment encourage students to learn about assessment criteria and ways of meeting these three ratings in essay writing. Even though researchers found that students are motivated to develop their writing skill through the rating experiences, they reported that self and peer marks tend to be different from the tutor marks in many ways.

Researchers also examined the factors affecting rating judgment. Cheng and Warren (1999) identified rating experience of the specific task as considerably contributing to the improvement of peer-ratings. Freeman (1995) commented that subjectivity leads to the

potential inconsistency of peer-ratings. Therefore, he suggested that careful training should be given to students. Magin (2001) added that the degree to friendship and social relation may yield bias. Since peers do not like to criticize their partners in the face, peers tend to flatter the score of their partners (Falchikov, 1995).

Other research has shown benefits of peer-grading. Patri (2002) noted that students thoroughly understand the assessment criteria when they try to be accurate in their peer score. Brown (1996) advocated that self- and peer-assessment provide authentic and dynamic evaluation, not solely relying on unilateral evaluation from the tutor. Topping and Smith (2000) described that learners can improve the quality of subsequent writing through the subjective feedback combined with assessment. Most of all, peer-assessment as a form of innovative assessment is susceptible to learner's needs and aims to enhance the quality of learning (Brown, Rust, & Gibbs, 1994).

In the language testing area, the concurrent validity of peer-assessment has been investigated for oral skills (Hughes & Large, 1993; Patri, 2002) and written skills (Kim, 2002). Patri (2002) found that peer-assessment is sufficiently comparable with teacher assessment in the case that assessment criteria are clearly described. Of few L2 writing test studies, J. T. Kim (2003) looked at the process and outcomes of peer-assessment in a placement test. He reported that learners tend to inflate their scores both in self-assessment and peer-assessment. However, learners showed high acceptance of peers' evaluation.

Overall, studies about the peer-rating suggested that the peer-grading is an innovative form of learning in classroom situation. However, there were still unkown about the peer-grading in the second language writing context.

III. METHOD

1. Research Question

Does collaborative rating really trigger psychometric unreliability in an English essay-writing test?

For this question, this study will compare and contrast the results on the three scoring modalities: expert, peer, and self. That is, the current research concentrates on the rater variance in order to examine the comparability of self- and peer-ratings with the expert ratings.

2. Subjects

104 students were solicited from the Korean college-level students who had studied

English language in Korea at least for the last 6 years. The subjects were requested to write an argumentative English essay in their home. Then, they were asked to perform ratings of three sample essays in a rater training session, to rate their own and a peer's essay, and to fill in a questionnaire in the classroom at the day of data collection. The participants at their class schedules were allowed to select one of the two days: November 2 and November 3, 2004. Each day, a brief oral description about the main research was made in the classroom.

91 (88%) of 104 participants majored in English language education and/or literature, while only a few students specialized in other fields. There were several reasons on restricting the subjects to those whose major was English related. First, the participants in that major were expected to have enough writing proficiency to make productive assessment. Since they had studied English grammar at least over 6 years, they were able to comprehend a variety of English sentences. Second, the subjects were selected because they were not only English language learners but also potential English language teachers in Korea. The participants were exposed to the same EFL learning context as other Korean learners. In other words, they shared the common sense of frustration and hardship with other college students in learning English. At the same time, they would be disciplined to evaluate student performance as a prospective English teacher or professional. Therefore, they might feel the sense of responsibility for educating English learners of post-secondary schools. For these reasons, rating tasks were best suited for subjects whose major is English related.

3. Instrumentation

1) Reading Article

Students were asked to write an English essay with the topic that a reading article dealt with. The article was carefully selected by the following considerations. First, it should include both positive and negative viewpoints toward a chosen topic. Second, syntactic structure and vocabulary used in the article were so easy that students clearly understood the underlying message. Third, the topic was more general than field-specific. By these considerations, the researcher chose a social topic, the journalism and privacy.

2) Scoring Benchmark

The benchmark for analytic rating was developed mainly with reference to the one developed by Willard-Traub, Decker, Reed, and Johnston (1999), TOEFL writing scoring

guide (ETS, 2000), the present UIUC ESL Placement Test (EPT)[1] feature analysis form, MATE rubric[2] (1998), and participants' suggestions.

The researcher accepted Willard-Traub *et al.*'s guideline largely in a content feature since it suggested comprehensive descriptors for assessing an ESL essay for academic purposes. MATE rubric was also considered in the scoring guide because it was designed with the empirical analyses of Korean EFL learners' writing performance. According to MATE research report on its writing test (2004), most Korean students were in intermediate level. No college-level examinees were ranked at the rudimentary or the expert level. Therefore, the benchmark for the main study adapted more descriptors corresponding to the intermediate level than the benchmark of the current EPT.

In designing a scoring benchmark, each of four features (organization, content, language use, and holistic) contained more than three semi-features. The purpose of the semi-feature was to make a clear-cut boundary between each feature. By rating such a semi-feature, a student was expected to rate an essay thoroughly and substantially, thus, understanding what a feature intended to measure in detail.

The semi-features were mainly adapted from the feature analysis form used in the current EPT. On the feature analysis form, there were 16 semi-features believed to calibrate the quality of an essay exhaustively. After some features were rewritten, deleted, and merged into the other features, the scoring benchmark for the current study included 12 semi-features: 3 (intro, body, and conclusion; cohesion; coherence) on organization, 5 (topic observance; support; writer's own idea; repetition; plagiarism) on content, and 4 (grammar accuracy; spelling; appropriate word choice; syntactic structure) on language use. Each semi-feature provided a descriptor that indicated the level of quality of the text. Every feature, then, was arranged by a five-step score category from 1 (lowest) to 5 (highest). The benchmark was printed in Korean and English to help some students clearly understand the descriptors.

3) Assessment Form

Two assessment forms were developed for the rating tasks. One was a self-assessment sheet in which a student would mark the first self-score and the second self-score, and the other a peer-assessment form in which he/she would assess the essay of his/her partner. Each sheet had columns in which students would defend their ratings in their mother

[1] The ESL Placement Test (EPT) is administered to foreign students, before the students start their first semester in the University of Illinois at Urbana Champaign. The purpose of EPT is to place students into a proper ESL level.

[2] MATE (Multimedia Assisted Test of English) is currently revising its rubric, referring to the analyses of a number of Korean EFL learners' speaking and writing scores.

tongue.

4. Procedure

An essay-writing assignment had given to subjects a week before the experiment was conducted. Each student was requested to write an English essay on a suggested topic, to include the content of a given article, to type his/her essay on the computer, and to submit two hard copies of the essay. Students were urged to print their student ID numbers. Even though they were allowed to use any resource relevant to the essay-writing tasks, they were strongly advised to write their own words and sentences. At the same time, they were encouraged to write at least one-page long English essays with the time limitation as they would experience in the TWE of TOEFL. Students returned two copies of their essays to the researcher three days before the experiment was conducted.

A rater training session was held on the day of the experiment. During the rater training, a scoring guide and three essay samples with rating boxes were distributed to subjects. As a first step, a proctor randomly formed groups-two students for each group. Then, the proctor handed out the scoring benchmark for holistic and analytic rating. The students were guided to carefully read through the scoring benchmark printed in their first language within 10 minutes. Each group then had oral discussion about the assessment criteria. If requested, the proctor provided clear instruction about the scoring criterion. After 5 minute's discussion was over, three writing samples were shown to the students. One writing sample simulated a low intermediate level's essay. The other writing sample simulated a high intermediate level's essay. After the students had completed scoring each essay sample independently and individually, each group then negotiated each other for the accurate rating of the sample essay. Finally, the proctor revealed the set of scores pre-determined by the expert raters whenever subjects finished negotiating scores on each essay sample. During the rater training session, students were encouraged to use their first language in oral discussion. Also, the proctor gave instruction in Korean.

Right after the rater training session was over, each subject was given a photocopy of his/her own essay, a photocopy of a peers' essay, and three assessment sheets: self-assessment sheet I, peer-assessment sheet, and self-assessment sheet II. The students were encouraged to assess their own essays and peer's essays individually and confidentially. Every subject was also guided to write comments justifying the scores on his/her own essay and his/her peer's essay. After the assessment tasks had been done, students were asked to change their seats according to the seating charts. Each student was seated with the partner predetermined by the researcher and, by this time, knew who was his/her partner. Then, students were strongly told that the peer score was not a result of judgment but a sort of diagnostic information. The students were also urged not to use offensive

words during peer-assessment. After the students listened to such the guidelines, each student in a group exchanged his/her score and the written comment on partner's essay, provided oral feedback, and negotiated the proper score for each essay. After the group discussion had been finished, students filled out their self-scores in self-assessment sheet II. The students were told that they were able to keep or correct their first self-scores upon their own decisions.

Two expert raters were invited to rate students' essays. They were selected because they successfully finished the rater training program set by a testing institution, because they had a MA or an equivalent degrees in TESOL program, and because they rated more than 100 Korean EFL learners' essays. The two raters are currently working as the official raters of MATE. In advance of the data collection, the researcher had one hour's meeting with each of two expert raters twice in order to make them clearly understand the benchmark. As soon as experiment had been done, the researcher sent the students' word-processed essays to expert raters via mail. The expert raters gave ratings on the same features (organization, content, language use, and holistic part) as students did.

4. Data Analysis

The current study utilized three statistical techniques: descriptive statistics as a preliminary analysis, intraclass correlation as a macro-level analysis, and FACETS research as a micro-level analysis. That was, three techniques together will help detect the psychometric reliability and validity of the multiple assessment design in the L2 essay-writing test, by investigating the inter-relationship among the facets and the intra-characteristics within each facet.

Specifically, many-facet Rasch measurement as a micro-level analysis was used to investigate the rating severity of every rater within each rater group (the first self-rating, peer-rating, the second self-rating, and expert-rating) on the basis of individual ability and item difficulty. The many-facet Rasch analysis optimally fitted for the main research context because the analysis told the rating characteristics of every rater, because it informed of how many significant biases were found as a result of interaction effect, i.e. interaction of a rater by an analytic area, and because it graphically ordered the degree of ability for test-takers, the degree of severity for raters, and the degree of difficulty for analytic areas. The main study analyzed the data with the computer program, FACETS, commonly used for the Rasch measurement (Linacre & Wright, 1999).

The next step for FACETS analysis was to identify as many potential sources of variation affecting the measurement as possible. In the main study, 'person', 'feature', and 'rater' will be specified as three facets- testing factors as a set of similar conditions of measurement. The facet 'person' referred to the systemic variance to be measured. The

facet 'feature' denoted 4 conditions (3 analytic and 1 holistic feature). The facet 'rater' consisted of 5 conditions (two self-raters, one peer rater, two expert raters).

IV. RESULTS

1. Descriptive Statistics

1) Inter-Rater Reliability

Table 1 reports Intra-Class Correlations (ICC) for the purpose of investigating inter-rater reliability of analytic/holistic features. Based on the ICC values, organization produces the highest ICC indices (.47 to .81), while content yields the lowest ICC indices (.13 to .72). That is, raters' judgments on content seem most inconsistent. In addition, raters disagree more in language use (.33 to .76) than in holistic rating (.42 to .79).

TABLE 1
Intra-Class Correlations (Unbiased Reliability Estimate) for Holistic/Analytic Ratings

Rater group	1	2	3	4	5
		Organization			
1. Self 1	-	.47	.81	.61	.61
2. Peer		-	.63	.48	.49
3. Self 2			-	.63	.55
4. Korean expert				-	.74
5. Native expert					-
		Content			
1. Self 1	-	.39	.72	.24	.13
2. Peer		-	.50	.38	.25
3. Self 2			-	.36	.32
4. Korean expert				-	.69
5. Native expert					-
		Language use			
1. Self 1	-	.38	.76	.33	.35
2. Peer		-	.59	.41	.41
3. Self 2			-	.41	.38
4. Korean expert				-	.72
5. Native expert					-
		Holistic			
1. Self 1	-	.48	.79	.45	.49
2. Peer		-	.63	.50	.42
3. Self 2			-	.54	.53
4. Korean expert				-	.77
5. Native expert					-

N = 104 per each feature. All values are intra-class correlations.

As seen in Table 1, students do not produce ICC indices that are as high as the expert raters do. This can be seen by comparison of the self 1/peer correlations (.38 to .43) with the Korean/native expert correlations (.69 to .77). This suggests that while an expert rater with a great deal of rating experience is trained as a consistent grader, students have no or very limited opportunities to rate an academic essay. The inter-rater reliability coefficients for content and language use are very low between the first self-rating and each of the expert ratings. This can be seen from the fact that for the self1-expert pairs, the ICC values range from .13 to .24 in content and from .33 to .35 in language use. The second self-rating, however, shows better consistency with the expert rating in content (.32 to .36) and language use (.38 to .41). This indicates that students adjust their self-scores after they learn of peer-grades.

2) Mean Scores

Table 2 suggests the mean scores of holistic/analytic features for each of the rater groups. Notable mean differences between self 1 (2.39 to 2.79) and peer (2.64 to 3.10) suggest that students initially underscore their own performance, whereas they overestimate their partners' essays to a great extent. The mean range (2.42 to 2.79) of self 2 indicates that after the students receive the feedback from their peers, then they relocate the initial self-scores to a certain point of the range between the first self-score and the peer score. In addition, a Korean expert's mean scores, 2.21 to 2.50, are closer to the first self-scores, 2.39 to 2.79, while a native expert's mean scores, 2.59 to 3.11, are near the peer-scores, 2.64 to 3.10. Overall, the second self-score acts like a middle point, an average of all types of ratings.

Regarding analytic/holistic features, the means of organization, 2.27 to 2.64, are lower than those of any other analytic areas. On the other hand, language use exhibits the highest mean scores, 2.50 to 3.11. This implies that rater groups show the most generous rating behaviors in language use and/or that students perform language use well.

TABLE 2
Mean Scores of Holistic/Analytic Features for Each of the Rater Groups

Features	Self1		Peer		Self2		KEX[a]		NEX[b]	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Organization	2.39	.94	2.64	1.01	2.41	.94	2.27	.99	2.59	.85
Content	2.62	.84	3.00	.88	2.72	.86	2.21	.77	3.05	.78
Language use	2.79	.94	3.10	.97	2.79	.91	2.50	.82	3.11	.76
Holistic	2.50	.80	2.95	.84	2.64	.86	2.25	.81	2.78	.82

N = 104 for each of the rater groups. [a] KEX = A Korean expert. [b] NEX = A native expert.

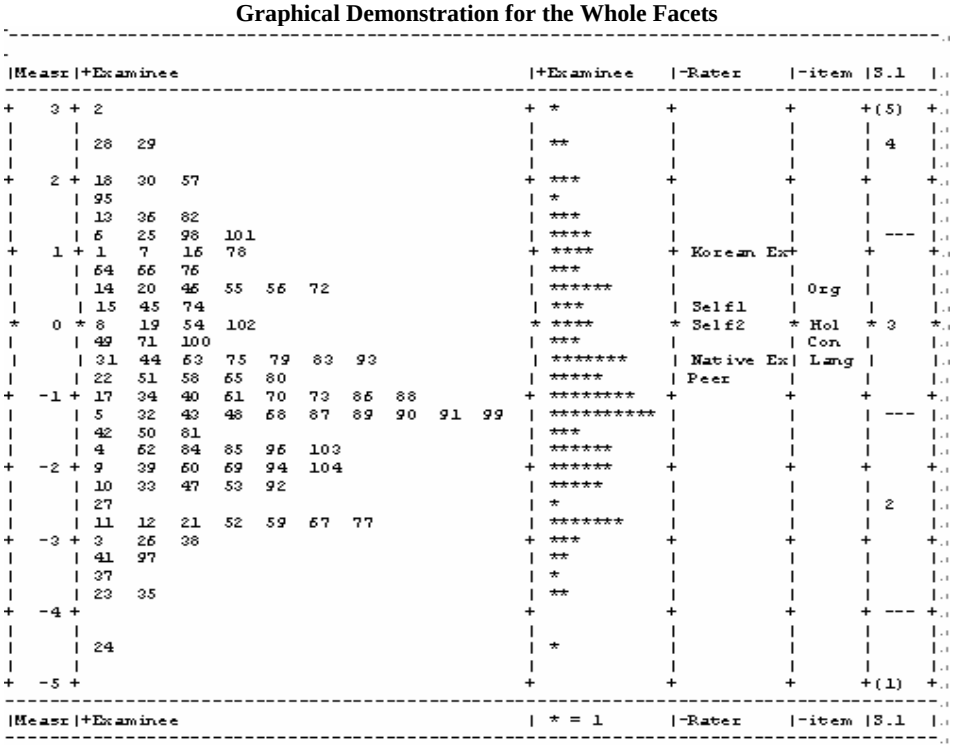
2. FACETS Ananlyses

The many-facet Rasch measurement analyses are conducted with FACETS, a computer program which converts ordinal data into interval scale. In this section, three facets are examined: features, raters, and persons. Each of the facets is separately analyzed based on the data derived from 2080 measurable responses (104 people x 5 rater groups per person x 4 writing features). A misfit analysis is also conducted to enlighten the details of the misfitting students whose score patterns are idiosyncratic. In addition, interaction analyses including each of the pairs (person by rater, person by feature, and rater by feature) are conducted in order to detect biases.

1) Total Facet Analyses

FACETS analyses provide a graphical contour of all facets on a yardstick. The measure on the left column of the yardstick represents the range of the logit scale, which indicates the ability measure for a person facet, the difficulty measure for a feature facet, and the severity measure for a rater facet. All these facets are positioned on the same logistic vertical ruler with every element marked on a single frame of reference. In every facet, an

FIGURE 1
Graphical Demonstration for the Whole Facets



element with the highest positive logit is located in the top, while an element with the lowest negative logit is in the bottom. The positive logit connotes the high ability for a person facet, the harsh rating for a rater facet, and the difficulty for a feature facet. Conversely, the negative logit denotes less ability for a person, generous rating for a rater, and ease for a feature. Finally, the farthest right column of the figure denotes the range of the logit each raw score category falls on.

Figure 1 graphically demonstrates the relative standing of all facets. The interval from the highest proficient student (student 2 with logit 3) to the lowest proficient (student 24 with logit -4.5) is approximately 7.5. This indicates a wide spectrum of student writing abilities. More than half of the students rank below the zero logit, which suggests that the distribution of students is positively skewed. Organization is the most difficult feature for students, while language use is the least difficult. The Korean expert is prominently harsher than any other raters. On the other hand, the peer-raters are most lenient.

2) Feature Analyses

Table 3 reports the statistical indices in terms of a feature facet. The most difficult writing feature is organization and the least difficult area is language use. The significant chi-square value of a feature facet, 110.9 with df = 3, suggests that the features have varying degrees of difficulty. Moreover, the separation index, 5.18, verifies a wide range of difficulty across the features. The high reliability index of the feature separation indicates that the difficulty of each feature is truly different from one another.

TABLE 3
Feature Measurement Report

Feature	Logit	*SE*	Fair-*M* average	Infit *MS*	Infit *z*
Language use	-.49	.07	2.85	1.1	1
Content	-.14	.07	2.71	1.1	1
Holistic	.10	.07	2.61	.7	-5
Organization	.53	.07	2.44	1.1	1
Mean	.00	.07	2.65	1.0	-.2

Observed count per feature = 520, Separation = 5.18, Reliability = .96, Fixed (all same) chi-square = 110.9 (df = 3, p = .00).

Given the infit indices, misfitting features (the infit mean-square index greater than 1.2 and/or infit z greater than 2) are not found. According to McNamara (1996), a misfitting feature does not allow people to predict person ability in a specific feature. Such a misfitting feature measures another trait that the other relevant features do not account for. In this sense, no misfitting feature indicates that all features together unidimensionally

measure one trait in the current test context. Notwithstanding, holistic judgment, called overfitting feature, is of primary concern both because its infit *z*, -5, is strongly negative and because its mean-square is less than .8. This means that the holistic component may be a redundant feature in predicting writing performance, or that students strictly control variability in order to avoid an extreme marking.

3) Rater Analyses

Tables 4 to 8 provide a summary of selected statistics on the rater facet. The rater facet analyses demand a different approach to interpreting separation index and reliability of the separation. In the ideal psychometric sense of inter-rater reliability, the separation index should be 1, which indicates that all raters reach perfect agreement. Therefore, a low reliability of separation is sometimes expected, in that the varying patterns of raters are controlled as being equally severe.

Another useful tip for the rater facet analyses is to compare observed agreement to expected agreement rate. The observed agreement percentage is the ratio of observed total ratings by a pair of groups to observed cases of agreements by the rater pair. The expected agreement as chance agreement is acquired from the marginal frequencies of the contingency table. If the observed agreement rate is too low in comparison with expected agreement rate, a model for predicting agreement is problematic. On the other hand, when the observed percentage of exact agreement is too much greater than the expected percentage, there is a possibility for raters to manufacture agreement by external factors. In that case, a rater possibly does not perform ratings independently (Linacre, 1989).

(1) For Total Feature

Table 4 denotes the analyses of the rater facet in terms of total analytic/holistic features. In the second column, the severity span between the most generous peer, -.66, and the least generous Korean expert, .94, is 1.60 logit. Students are more severe in the self-rating than in the peer-rating. Given the separation index, 7.27, and the highly significant χ, 265.4, all raters demonstrate varying degrees of severity with a reliability index of .98. In other words, raters consistently differ from one another in overall severity. According to the exact agreement rate displayed in the seventh column, the second self-rating is highly dependent upon the other type of ratings.

Looking at the fit statistics, the peer group is identified as misfitting. Fit value for the peer group is beyond the range of two standard deviation around the mean. The peers' rating pattern is not predictable, indicating that the peer group is not intra-consistent. This calls for a need of rater training for minimizing the misbehavior of peer-raters.

TABLE 4
Rater Measurement Report for Total Feature

Rater	Logit	*SE*	Fair-*M* Average	Infit *MS*	Infit *z*	Exact agreement *p* Observed	Exact agreement *p* Expected
Peer	-.66	.08	2.92	1.2	2	38.9	41.0
NEX	-.55	.08	2.87	.9	-2	39.8	41.7
Self 2	.04	.08	2.63	.9	-1	51.3	43.3
Self 1	.23	.08	2.56	1.1	1	47.0	43.1
KEX	.94	.08	2.28	1.0	0	39.8	39.4
Mean	.00	.08	2.65	1.0	-.1	43.4	41.7

Observed count per rater = 104, Separation = 7.27, Reliability = .98, Fixed (all same) chi-square = 265.4 (*df* = 4, p = .00).

(2) For Organization Feature

Table 5 displays rater measurement in terms of organization. The peer-rater group, -.64, is more generous than any other rater groups, whereas the Korean expert, .68, is least lenient. Both the separation index, 2.39, and the reliability of separation indices, .85, corroborate that the varying degrees of severity exist among rater groups in terms of organization rating. Compared to the total feature, organization ratings have a restricted range of rater severity (from -.64 to .68). This suggests that raters control intra-rater variability to some degree in organization rating. On the other hand, the relative order of severity along the rater groups does not differ from total to organization feature. Fit statistics show that the peer group conducts the misfitting rating behavior, while the second self-group does not allow enough variance for discriminating one student from another.

TABLE 5
Rater Measurement Report for Organization

Rater	Logit	*SE*	Fair-*M* average	Infit *MS*	Infit *z*	Exact agreement *p* Observed	Exact agreement *p* Expected
Peer	-.64	.18	2.68	1.5	3	40.4	48.9
NEX	-.44	.18	2.61	.9	0	44.0	49.8
Self 2	.17	.18	2.42	.7	-2	54.1	50.8
Self 1	.23	.18	2.40	.9	0	49.8	50.7
KEX	.68	.19	2.25	1.0	0	45.4	49.1
Mean	.00	.18	2.47	1.0	-.2	46.7	49.9

Observed count per rater = 104, Separation = 2.39, Reliability = .85, Fixed (all same) chi-square = 33.5 (*df* = 4, p = .00).

(3) For Content Feature

Table 6 suggests the rater measurement in content rating. The native expert rater is most lenient given the logits, -.99. By contrast, the Korean expert rater is harshest by the logits, 1.61. Since the logit value, 1.61, is far deviated from the group of the other logits (-.99

to .30), the Korean rater is conspicuously strict in content. A wide logit interval (2.60) from the native to the Korean expert, a great separation index (5.37), and a high reliability value (.97) confirm to the varying degrees of severity across the raters. Like the other features described above, content rating reliably separates raters into different levels of severity. Infit mean-square values range from 0.8 to 1.2 across all raters. This suggests that no raters show idiosyncratic patterns in content ratings. In other words, they are self-consistent in content ratings.

TABLE 6
Rater Measurement Report for Content

Rater	Logit	*SE*	Fair-*M* Average	Infit *MS*	Infit *z*	Exact agreement *p*	
						Observed	Expected
NEX	-.99	.17	3.02	1.1	0	33.2	41.4
Peer	-.86	.17	2.97	1.1	0	35.6	42.4
Self 2	-.06	.17	2.69	.8	-1	47.4	45.5
Self 1	.30	.17	2.57	1.1	0	43.3	45.3
KEX	1.61	.18	2.17	.9	0	33.9	36.7
Mean	.00	.17	2.68	1.0	0	38.7	42.3

Observed count per rater = 104, Separation = 5.37, Reliability = .97, Fixed (all same) chi-square = 143.6 (*df* = 4, p = .00).

(4) For Language Use Feature

Table 7 delineates rater measurement for language use. The native expert (-.99) in the top row is more generous than the Korean expert (1.61) in the bottom row by logit 1.82. Separation index, 3.85, and reliability of the separation, .94, evidence that the rater groups' severity is consistently different from each other. Given the accepted range of infit *MS* and infit *z*, every rater is identified as infitting.

TABLE 7
Rater Measurement Report for Language Use

Rater	Logit	*SE*	Fair-*M* Average	Infit *MS*	Infit *z*	Exact agreement *p*	
						Observed	Expected
NEX	-.74	.17	3.11	.9	0	38.2	44.0
Peer	-.71	.17	3.10	1.2	1	41.1	44.2
Self 2	.18	.17	2.80	.8	-1	51.7	46.5
Self 1	.20	.17	2.79	1.1	0	46.2	46.4
KEX	1.08	.17	2.48	1.0	0	35.8	41.8
Mean	.00	.17	2.85	1.0	-.1	42.6	44.6

Observed count per rater = 104, Separation = 3.85, Reliability = .94, Fixed (all same) chi-square = 78.8 (df = 4, p = .00).

(5) For Holistic Component

Table 8 reports rater measurement in terms of holistic ratings. The distribution of holistic rating from the most lenient peer group, -1.37, to the harshest Korean expert rater, 1.58, is wide. According to the separation index, 4.92, and the reliability of separation, .96, significant difference among rater groups is consistent. That is, the varying degrees of severity are obvious throughout the rater groups. Regarding intra-rater consistency, the infit statistics (.8 to 1.2) suggest that every rater group produces the predictable rating pattern. That is, all rater groups are self-consistent in holistic rating.

TABLE 8

Rater Measurement Report for Holistic Component

Rater	Logit	*SE*	Fair-*M* Average	Infit *MS*	Infit *z*	Exact agreement *p*	
						Observed	Expected
Peer	-1.37	.20	2.95	1.2	1	38.5	46.0
NEX	-.63	.20	2.77	1.0	0	43.8	51.2
Self 2	-.09	.20	2.63	.8	-1	52.2	52.7
Self 1	.51	.20	2.46	1.0	0	48.8	52.0
KEX	1.58	.21	2.19	1.0	0	44.2	45.1
Mean	.00	.20	2.65	1.0	.0	45.5	49.4

Observed count per rater = 104, Separation = 4.92, Reliability = .96, Fixed (all same) chi-square = 123.3 (df = 4, p = .00).

4) Person Analyses

Table 9 provides a summary of person measurement for 104 examinees. The mean ability of persons is -0.80 logit with a standard deviation of 1.52. The examinee ability ranges from –4.38 to +2.89 logit. The widespread interval means that the essay-writing ability varies across the sample examinees. The ability difference is supported by the large separation index, high reliability of the examinee separation, and great chi-square values. The separation index, 4.24, connotes that the observed variance in measurement of examinee ability is about four times the measurement error. The high reliability statistics, .95, suggest that this analysis reliably discriminates the differing abilities from each other.

TABLE 9

Person Measurement Summary

Statistic	Logit	*SE*	Fair-*M* Average	Infit *MS*	Infit *z*
M	-.80	.36	2.66	1.0	-.2
SD	1.56	.01	.62	.5	1.5

Observed count per person = 20, Separation = 4.23, Reliability = .95, Fixed (all same) chi-square = 1896.2 (*df* = 103, p = .00).

5) Misfit Analyses

Misfit analyses identify seven misfitting students who fail to demonstrate a consistent behavior pattern across the features and across the raters. Four of the seven misfitting students give relatively high self-scores in comparison with experts' scores, which contribute to the students' unpredictable and unexpected behavior pattern. Conversely, three students produce self-scores much lower than the expert ratings.

Peer-ratings create varying effects on the rating pattern of the students. For three misfitting students, a peer-rating reinforces the self-overestimation. For four students, the peer-rating better predicts the experts' scores than the self-rating does. In general, although peer-rating sometimes prevents the severe self-underestimate, it also bolsters self-inflation.

6) Bias Analyses

The magnitude of the interaction is also calibrated by interaction logit and z-score. Like previous facet analyses, a plus/minus sign of an interaction logit indicates a direction of a response pattern. If a logit is -3.05 in a person-by-rater interaction, the rater's rating is too generous for the person in comparison with the other raters. A z-score determines the significance of an interaction pair in the bias analyses. An interaction with an absolute z-score greater than 2 is categorized as a significant bias. In this way, the bias analysis identifies an individual element that consistently responds to another facet in a way that is different from the other elements. The current study does not include person-by-feature interaction since the main focus is on students' rating performance. All interaction pairs include a rater facet in the current study.

(1) Person-by-Rater

There are potentially 520 (104 subjects X 5 rater groups) bias pairs in terms of person by rater interaction. Of these, 63 interactions, or 12.1% of the total 520 bias terms, demonstrate significant patterns. The frequency difference between the most significant bias group, sixteen peer-raters, and the least significant bias group, nine self 2 raters, is only 7. This indicates that the number of significant interaction does not significantly differ through rater groups.

(2) Rater-by-Feature

Table 10 provides a bias/interaction report for the rater-by-feature interaction. Of the potential 20 bias pairs, only two are found to have significant biases. Moreover, only a Korean expert rater commits significant interaction with two of the features: organization and content. In the KEX-by-organization interaction, a large z-score, -2.67, and mean-

square value, 1.2, suggest that the Korean expert consistently demonstrates a more generous rating behavior in organization than she does in the other features. On the other hand, the Korean expert remains strongly more severe than expected in the content rating.

TABLE 10
Bias/Interaction Report for the Rater and Feature

	Rater		Feature	Interaction			
Raters	Logit	Features	logit	Logit	*SE*	*Z*	*MS*
KEX	.94	Organization	.53	-.43	.16	-2.67	1.2
KEX	.94	Content	-.14	.40	.16	2.45	.9

There are empirically 20 bias pairs.

V. CONCLUSION

1. Summary

Quantitative techniques such as descriptive statistics, Intra-Class Correlation (ICC), and FACETS analyses the show evidence that a peer rater group demonstrates a generous rating behavior to a great extent. The overgenerous rating pattern causes the misfitting rating, which disallows us to predict an operative pattern from the peers' rating data. FACETS analyses indicate that only a peer-rater group commits unreliable patterns. This suggests the need of a rater training program for the peer-rater group. On the other hand, self-raters show relatively severe ratings at the beginning stage. After the self-raters access the peer- feedback, they position their essay scores higher than the previous self-ratings. The ICC indices suggest that students do not quite fairly agree with the expert raters in content and language use, while the students and the experts show relatively high agreement in organization and holistic ratings. According to descriptive statistics, all rating groups except the Korean expert display the agreement in the relative standing of feature difficulty in their ratings. That is, they perceive that organization is most difficult, and respectively, holistic score, content, and language use. Overall, the quantitative analyses of the current study propose conflicting results about the research question. Even though the descriptive statistics propose that less than 50% of total rating cases reaches agreement between students and each of two experts across all the features, the micro-level analysis using FACETS evidences that every rater group produces a common rating pattern except the peer rater group. This suggests that while the self-ratings are reliable, the peer-ratings are not comparable with the other types of ratings.

2. Implications and Suggestions

Since the data are collected from the Korean EFL learners residing, the current study suggests a lot of practical and applicable findings generalized to EFL contexts.

First, the collaborative rating procedure is recommended for the students who major in education. The students of the educational fields are necessarily trained to be teachers. Each of the students is required to practice a rating, which is one of the tasks that he/she as a prospective teacher will perform in their future school. In particular, the heads or faculties in English Language Education (ELE) may introduce the collaborative rating procedure into the curriculum in order to provide the motifs to the students in the major. That is, a language learning course that links the teacher education may better motivate ELE students to involve in the class activity.

Second, it is suggested to develop a rater training program specialized for the students. Since a rater training session enhances students' competence and confidence in the rating skill, a step-by-step guideline needs to be established for the development of the rater training program. In order to customize the rater training to students' demands, students' feedback should be collected through all developmental stages of the rater training program. The students, for example, are able to participate in the revision of the score rubric. Students' responses to the survey are used to identify a feature in which the students show unreliable rating pattern. In doing so, students' involvements in the rater training developmental process encourage them to understand what the rating is, thereby inviting high motivation.

During the rater training session, some activities should be taken to reduce the overrating cases by the peer-raters. It is recommended to introduce the following activities: name concealment, explicit instruction, descriptor clarity, exposition of expert's score, and overall evaluation. In the name concealment, the partner's name should be unknown to the students in order to prevent them from conducting possible biases. The explicit instruction refers to a persuasive and appealing suggestion. In the explicit instruction level, students should be informed of the fact that a rating involves the commitment to honestly measure the peer students' performance. Student should also be aware of the fact that a score is not a judgment of a peer's outcome but a means of helping the peer's learning. For the purpose, an instructor needs to provide a friendly and cooperative atmosphere for reducing affective filter. The descriptor clarity involves the process in which students are provided with the clear explanation about each score descriptor. The exposition of the expert scores on the sample essay allows students to see their scores more objectively. In the overall evaluation, students monitor their own rating patterns in a reflective manner. The students may be encouraged to contemplate what consequences their ratings may result in.

Third, it is demanded to design curriculums/programs that train the students as expert

raters in the EFL context. This means that students should be prepared for a large-scale writing test. For the purpose, students need to be under the context in which they are obliged to rate their classmates. Every student, for example, is allowed to participate in one of the evaluation committee groups, each of which consists of two students and an instructor. Each committee determines the grading of each anonymous peer's academic achievement in the course. In this way, students are provided with moderate tension that functions as an impetus of the rating practice. The students also are expected to feel the sense of responsibility for the peer's learning, which is a canonical virtue in the education.

Finally, the author would like to point out that the research agenda about students' involvements in the rating process is rarely discussed among the Korean EFL professionals let alone among ESL professionals. Therefore, the author is not definitely certain that the above suggestions are the well-timed and suitable treatments for the collaborative rating context. These suggestions need to be confirmed and revised by future empirical studies.

People consciously and unconsciously believe that a rating should be the objective measurement of something with an illusion that accuracy is an ontological foundation for the rating. That is, it is very common to connect psychometric values (objectivity and accuracy) to the concept of the rating. Therefore, only a well-trained expert has been empowered to conduct the rating in order to safeguard the psychometric values. Excessive focus on the accuracy, however, leads us not to be aware of the qualitative nature of the rating. The results of the multifacet analyses suggest that students can be reliable raters when they self-rate their own essays. The rating also encourages students to take the responsibility for their peers' learning. Nevertheless, should the rating remain in one of most conservative areas in which a student's participation is not welcomed?

Ultimately, the current study proposes that a rating task can be a form of alternative assessment. The author hopes that the findings of the present study initiate the extensive arguments about the learner-directed rating as a form of alternative assessment.

REFERENCES

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, *6*, 14-29.

Blanche, P., & Merino, B. (1989). Self-assessment of foreign language skills: Implications for teachers and researchers. *Language Learning*, *39*, 313-340.

Boud, D. (1991). Implementing student self-assessment. Sydney: HERDSA.

Brown, S. (1996). *Assessment*. Retrieved August 24, 2004, from Oxford Centre for Staff and Learning Development Web site: http://www.lau.ac.uk/deliberations/assessment/invite.html.

Brown, S., Rust, C., & Gibbs, G. (1994). Strategies for diversifying assessments in higher education. In *Involving students in the assessment process* (Ch. 5). Retrieved August 24, 2004, from Oxford Centre for Staff and Learning Development Web site: http://www.lgu.ac.uk/deliberations/ocsd-pubs/div-ass5.html.

Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment and Evaluation in Higher Education, 24*, 301-314.

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis. *Language Testing*, *2*, 164-179.

Ekbatani, G. (2000). Moving toward learner-directed assessment. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp. 1-11). NJ: Lawrence Erlbaum Associates.

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

Falchikov, N. (1995). Peer feedback marking: Developing peer-assessment. *Innovations in Education and Training International*, *32*, 175-187.

Freeman, M. (1995). Peer-assessment by groups of group work. *Assessment and Evaluation in Higher Education*, *20*, 289-299.

Hamp-Lyons, L., & Lynch, B. K. (1998). Perspectives on validity: a historical analysis of language testing conference abstracts. In A. J. Kunnan (Ed.), *Validation in Language Assessment* (pp. 253-276). Mahwah, NJ: Lawrence Erlbaum Associates.

Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7,* 172-198.

Hughes, I. E., & Large, B. J. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education*, *18*, 379-385.

Janssen-van Dieten, A. (1989). The development of a test of Dutch as a second language: The validity of self-assessments by inexperienced subjects. *Language Testing, 6*, 30-46.

Kim, Jungtae. (2003). *Computer-delivered ESL placement test at the University of Illinois*. Unpublished master's thesis. Illinois: University of Illinois at Urbana-Champaign.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly*, *19*, 73-87.

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

Linacre, J., & Wright, B. (1999). *FACETS,* Version 3.17 [Computer program]. Chicago: MESA Press.

Littlwood, W. (1996). Autonomy: An anatomy and a framework. *System*, *24*, 427-435.

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

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing*, *18*, 351-372.

MacIntyre, P. D., Noels, K. A., & Clement, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*, 265-287.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer-assessment of group work. *Studies in Higher Education*, *26*, 53-63.

Magin, D., & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: How reliable are they? *Studies in Higher Education*, *26*, 287-298.

MATE. (1998). *Rater training guide book.* Seoul: Sookmyung Women's University Press.

MATE. (2004). Retrieved October 10, 2000, from http://www.mate.or.kr/research_04.html.

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

Mowl, G., & Pain, R. (1995). Using self and peer-assessment to improve students' essay writing: A case study from geography. *Innovations in Education and Training International, 32*, 324-335.

North, B. (2000). Defining a flexible common measurement scale: Descriptors for self and teacher assessment. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp. 12-47). NJ: Lawrence Erlbaum Associates.

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

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, *19*, 109-131.

Pierce, B. M., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of control. *Applied Linguistics*, *14*, 25-42.

Rivers, W. P. (2001). Autonomy at all costs: An ethnography of metacognitive self-assessment and self-management among experienced language learners. *The Modern Language Journal*, *85*, 279-290.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*, 1-20.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*, 4-14.

Shin, Dongil. (2003). *Hankukui Yeongeo Pyongahak 1: Chihumgaebal-pyun* [English Language Testing Series 1: Test Development]. Seoul: Hankuk Munwhasa Press.

Shohamy, E. (2001). *The power of tests*. Edinburgh: Longman.

Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, *19*, 69-75.

Strong-Krause, D. (2000). Exploring the effectiveness of self-assessment strategies in ESL placement. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp. 49-73). NJ: Lawrence Erlbaum Associates.

Topping, K. J., & Smith, E. F. (2000). Formative peer-assessment of academic writing

between postgraduate students. *Assessment and Evaluation in Higher Education, 25*, 149-169.

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*, 41-84.

Applicable levels: tertiary level
Key words: teacher education, self rating, peer-rating, second language writing test

Ho Lee, Ph. D.
Dept. of Educational Psychology
Second Langauge Acquisition and Teacher Education
Univeristy of Illinois at Urbana-Champaign
Tel: (02)938-4479
Email: kinghoya@naver.com

Received in May, 2005
Reviewed in June, 2005
Revised version received in August, 2005