Learner Involvement in Self- and Peer-assessment of Task-based Oral Performance*

Sang-Ki Lee & Sumi Chang
(University of Hawai'i at Mānoa)

Lee, Sang-Ki and Chang, Sumi. (2005). Learner involvement in self- and peer-assessment of task-based oral performance. *Language Research* 41(3), 711-735.

The new understanding of learner role as active participants in learning has accompanied a paradigm shift in assessment practices. Learners are perceived to be responsible not only for their own learning but also for the assessment of their performances in terms of its procedures and rationales (Cheng & Warren, 2005; Falchikov, 1986; Luoma & Tarnanen, 2003; Orsmond & Merry, 1996). This study explores the plausibility of employing self- and peer-assessment as alternative approaches to assessment of seven Korean learners' oral presentation task performance. In order to cope with the subjectivity aspect of evaluating behavior, the study encouraged learner involvement in task assessment, which was operationalized as students' participation in the development of assessment sub-criteria and learner training as well as discussion regarding the criteria. The results from three periods of students' self- and peer-assessment on their presentation performance revealed that learner involvement in assessment can lead to a high comparability among three different assessment types; self-, peer-, and teacher-assessment. The students also reported that they had positive attitudes toward the alternative assessment.

Key words: self-assessment, peer-assessment, task-based performance assessment, learner involvement, Korean learners

1. Introduction

In recent years, student-centered learning has become increasingly important in language learning (Luoma & Tarnanen, 2003; Taras, 2002). Cross (1996, as cited in Taras, 2002) lists student involvement and assess-

* We would like to thank three anonymous reviewers for their critical comments and suggestions. We would also like to thank John Norris for his invaluable support and input throughout the development of this paper.

ment as interdependent conditions of excellence in education. According to Taras (2002), increasing student involvement in their works, oral presentation, and task-based learning will further enhance student-centered learning. Task-based language learning calls for the learner to be involved in one's learning by actively using the target language. Task is regarded as being communicative with real-world uses to accomplish a specific purpose (Stanley, 2003). Task will be further discussed in section 3.2 of this paper.

Student-centered learning brings about active student involvement, and the students are expected to be independent and autonomous. However, there exists a contradiction between the aim of student-centered learning and traditional assessment (Taras, 2002). Conventional psychometric conceptions in educational testing have shortcomings that prevent it from being the most relevant tool of assessment. They are concerned with the precision of scores and do not provide the opportunity to use their newly-acquired knowledge for the learners (Moss, 2003; Wiggins, 1993a). In line with the educational aim of helping the individual learner become a competent intellectual performer rather than a passive selector of prefabricated answers, assessment practice also needs to undergo changes. As alternative approaches to assessment, self-assessment (SA) and peer-assessment (PA) have been actively discussed recently, but research results are conflicting. A small-scale it may be, it will be worthwhile to examine the effects of learner involvement as well as students' attitudes toward self- and peer-assessment of task-based oral performance.

1.1. Self- and Peer-assessment as Assessment Alternatives

A surge of interest in self-directed learning and learner autonomy has brought advances in self-assessment, where self-reflection by learners on their own performance is emphasized. Self-rating is an alternative paradigm where not only the teacher assesses learners, but learners share the ownership of rating (Louma & Tarnanen, 2003; Taras, 2002). Studies indicate that self-ratings can be valid measures of communicative language abilities (Bachman & Palmer, 1989; Heilenman, 1991). Use of SA is encouraged because learners are seen to develop responsibility and awareness for their own learning and performances (Boud, 1989; Heilenman, 1991; LeBlanc & Painchaud, 1985; Luoma & Tarnanen, 2003;

Orsmond, Merry, & Reiling, 1997; Patri, 2002).

In this vein, peer-assessment has also gained much attention especially as a means for enhancing the learner-centeredness in assessment (AlFallay, 2004; Cheng & Warren, 2005; Louma & Tarnanen, 2003; Orsmond & Merry, 1996; Patri, 2002; Stefani, 1998; Taras, 2001, 2002). Research shows that successful students engage in metacognitive activities to plan and monitor their learning and continually assess their skills as learners. Research on the evaluation of teaching and learning also suggests that students are reliable sources of information about the effects of teaching or its impact on their learning. In teaching, teachers typically undergo an implicit automatic process of gathering information from the students (e.g., comments, questions, non-verbal language), and they depend on their impressions of student learning to make important judgments. However, these informal assessments are rarely made explicit, nor tested, and they are not checked against the students' own impression or peer impression (Angelo & Cross, 1993). Valuable explicit components such as students' self- and peer-assessment should not be overlooked.

1.2. Suggestions for Objectivity of SA and PA

One caveat that has been extensively raised among researchers is the validity of SA and PA as assessment alternatives. Can students evaluate the performances of themselves and their peers accurately? This question reflects the common assumption among students as well as teachers that teacher-assessment (TA) is always correct, whereas SA and PA may not always be correct (Boud, 1989; Kwan & Leung, 1996; Orsmond & Merry, 1996; Orsmond et al., 1997; Patri, 2002; Stefani, 1994).

With the surge of interest in SA and PA as aforementioned, it should be noted that the results of empirical studies on the validity of SA and PA have been inconclusive. Based largely upon a significant correlation between SA and TA, Bachman and Palmer (1989), Williams (1992), Stefani (1994), and Oldfield and Macalpine (1995) argue that SA may be a valid and reliable measure. However, some other studies have failed in reaching the same conclusion (Hughes & Large, 1993; Orsmond et al., 1997). When it comes to the validity of PA, Hughes and Large (1993), Freeman (1995), and Cheng and Warren (2005) report a high accuracy of PA comparable to TA, while the results of Kwan and Leung (1996) and Orsmond et al. (1997) run counter to these positive observations. In addition, Patri (2002) studied the agreement between TA, SA, and PA with and without peer feedback. It was found that while the accuracy of PA could be improved with peer-feedback, SA was not so either in the presence or absence of peer feedback.

Some insightful suggestions have been made by many researchers in order to cope with the subjectivity aspect and to develop more reliable and valid SA and PA instruments. First, a number of researchers emphasize the importance of learner training for the accuracy of SA and PA (AlFallay, 2004; Cheng & Warren, 2005; Freeman, 1995; Kwan & Leung, 1996; Oldfield & Macalpine, 1995; Orsmond et al., 1997; Patri, 2002; Stefani, 1998; Taras, 2001, 2002). They argue that "students must be given adequate training and practice … in order to minimize potential inconsistencies associated with subjectivity" (Patri, 2002, p. 111). Second, peer-feedback may play a critical role in improving the objectivity of SA and PA. In the case of PA, in particular, Freeman (1995) and Patri (2002) found that peer-involvement and peer-discussion enable students to assess more accurately to a degree that PA is highly comparable to TA. Third, and the most importantly, students' active participation in defining assessment criteria has been encouraged (Cheng & Warren, 2005; Kwan & Leung, 1996; Orsmond & Merry, 1996; Orsmond et al., 1997; Patri, 2002; Stefani, 1994; Taras, 2001, 2002). Students should be provided with clear guidance and understanding of the assessment criteria by being involved in the procedure of setting up the assessment criteria. As Cheng and Warren (2005) put it, "it is important for both learners and teachers to be involved in and have control over the assessment methods, procedures, and outcomes, as well as their underlying rationales" (p. 93).

2. Purpose

The present study explores the feasibility of employing valid SA and PA as alternative approaches to assessment of Korean learners' oral presentation performance. The theoretical implications of the objectivity of SA and PA motivated the small-scale experiments of this study, which were specifically designed to address the following questions:

(1) What are the effects of learner involvement on the validity of self-assessment and peer-assessment?

(2) Is there a consistency in the students' attitude toward the three different types of assessment, i.e., self-assessment, peer-assessment, and teacher-assessment?

In making recommendations to help language teachers select, use, and evaluate language tests, Norris (2000, 2002) recommends to focus on assessment instead of just tests, to clarify the intended use of the test, and to evaluate the outcomes of assessment. Every language assessment must have clear intended use as a starting point. Who the test users are, what information the test should provide, why the test is being used, and what consequences the test should have are some questions that must be considered.

In this study, an assessment was designed to measure how well learners can use their knowledge of the Korean language in a formal oral performance. The assessment of the oral presentation task should provide information on the students' ability to do a presentation in Korean using the appropriately formal language. Various grammatical features of the Korean language (e.g., sentence enders, honorifics, deferent forms, etc.) determine the formality levels. For the learners of Korean as a foreign language, appropriate usages of these linguistic features are difficult (Sohn, 2001). Norris et al. (2002) noted that test instruments and procedures that are the best match with the assessment purposes should be used. Presentation was chosen as the target oral performance task because presentation requires use of formal language features, thus the best match to the purpose of the assessment. The test user is the teacher, who uses it for the purpose of assessing students' speaking ability and gives a grade. The assessment also provides an opportunity for the learners to reflect on their own as well as their peers' speaking abilities. Each presentation related to a chapter of the textbook, and the grade the students received on the presentation comprised 20% of the chapter test grade (i.e., Out of possible 100 points of each chapter test, a maximum possible points of 20 derived from the presentation grade).

3. Methods

3.1. Participants

The participants in the study were seven learners of Korean language enrolled in Korean 301, the fifth-semester Korean course at the University of Hawai'i at Mānoa (UHM). The students have been placed into the fifth-semester level class either through a placement test or by having successfully completed one of the fourth-semester Korean courses. Four students were heritage students (i.e., Korean-American students whose parents spoke Korean), while three students were non-heritage L1 speakers of English. The 16-week course met three times a week, fifty minutes each time. The participants ranged from freshman to senior at UHM, and five of the seven students were females and two were males.

3.2. Task

In assessment, it is performance on tasks in contextualized situations that can provide more comprehensive information on learners' capacity to effectively use and demonstrate the knowledge they have learned (Eisner, 1999; Haertel, 1999; Wiggins, 1993b). Tasks are the real-world things people do in everyday life (Long & Norris, 2000). Task-based language teaching (TBLT) attempts to utilize the benefits of *focus on meaning* and also adopts *focus on form* for increased accuracy. In this study, focus on meaning is shown in conveying the content of presentations to the audience. Focus on form involves drawing students' attention to linguistic elements of language in context, and it is done through the assessment of the Language component in presentations.

Korean 301 covers the four skills of listening, speaking, reading, and writing. Compared to cognate languages of English, Korean might be considered one of the most difficult languages for native speakers of English to learn in part because of its intricate hierarchical system of honorifics and formality. In the third-year Korean courses at UHM, materials are presented to help the learners achieve high levels of proficiency in interpersonal as well as interpretive and presentational communications (Sohn & Lee, 2003). One of the course goals is being able to use formal language appropriately.

The students were given the task of making three 3- to 5-minute oral presentations to the class. The presentation topics were the titles of the textbook chapter being covered at the point in time in semester: "Korean cultural assets" for presentation number 1 (P1), "book report" for presentation number 2 (P2), and "Korean etiquette or customs" for presentation number 3 (P3). For each presentation session, the students narrowed down the chapter theme to a more detailed topic (e.g., Pulkuk Temple, living cultural treasures, etc. for P1), did some research, and prepared an outline. They were encouraged to use vocabulary, grammar points, and expressions covered throughout the semester, as the presentations began at week 10 of the 16-week semester, and ended at week 14. The student presentations were videotape recorded.

3.3. Assessment Procedure

3.3.1. Presentation Number One

During P1, the students and the teachers were asked to write their comments on each presenter in an open-ended way in sentences or in phrases on the given form with blank lines (Appendix A). The students were informed that peer evaluation would count towards the evaluator's grade and not towards the presenter's grade. In other words, when Student A is presenting and Student B is evaluating Student A, the thoroughness of the evaluation was to be graded and taken into consideration of Student B's grade. For PA, the students were asked to look for four areas of COLD which are Content (C.), Organization (O.), Language (L.), and Delivery (D.), but detailed explanation of each area was not provided. After each presentation, a few minutes were allowed to finish writing PA and TA. Then, as homework, each student watched the video recording of her/his own presentation and wrote an open-ended SA using the same open-ended format as the one used for PA in class. Throughout the weeks of presentation process, the students were frequently reminded of their ability to assess and the underlying rationale for doing so (e.g., awareness-raising). When the class met for the following class period, the PAs and TAs were returned to the students for them to review. Based on the comments received on all three assessments during P1, assessment criteria were developed for use in P2 and P3 (see the RESULTS section for details). The self- and peer-assessment for P1 was conducted without in-depth discussion or student involvement in the assessment criteria.

3.3.2. Presentation Number Two

Two weeks after P1, the students did P2. Prior to the presentation session, the students were given the new assessment form to examine and become familiar with the format. The form contained sub-criteria of

COLD with Likert scale for rating presentation quality (Appendix B). The same presentation and assessment procedures were followed as in P1, but using the new assessment form. For homework, the students were instructed to first do SA using the new self-assessment form, and then to fill out a post-presentation survey to gauge their comfort level with SA and PA and also to see how much they were responsible in conducting the assessments (Appendix C).

3.3.3. Training Sessions

During the three class periods preceding the P3 session, the instructor provided training and discussion opportunities for the students. After some general comments about the oral presentations, each sub-criteria of COLD was explained using a worksheet. The students seemed to clearly understand the detailed criteria, with an exception of item number 3 under the Delivery category which is "Rapport with and sensitivity to audience." The item was explained as being aware of whether the audience understands the presentation. Two video segments were shown as model examples to demonstrate clear introduction, use of transition words, appropriate sentence endings, accurate grammar, etc. Finally, the students were given another worksheet on common mistakes in grammar and expressions that were displayed during the first two presentation sessions. The training sessions lasted about 15-20 minutes for each of the three sessions, totaling 55 minutes.

3.3.4. Presentation Number Three

Two weeks after P2, P3 was conducted following the same procedure and format as P2 described above. In addition to the post-presentation survey, an in-depth interview was conducted with each student in order to examine their attitudes toward the presentation task performance and the evaluation experiences. The P3 assessment was conducted after student discussion and involvement in criteria revision, and will allow examination of the effects of learner involvement and any changes in the student attitude.

4. Results

4.1. P1 Assessment

The comments from the P1 assessment were examined and categorized to form sub-criteria for each of COLD. First, for PA there were 42 entries, each of the seven students providing PA for six of their peers. The students made comments mostly on Content and Delivery. There were all 42 entries for Content, mentioning general comments, interest level of the topic, informativeness, or clarity. Forty entries were made about Delivery, and almost everyone mentioned eye contact and comfort/confidence level. In an effort to help the students feel more at ease for the P1 assessment, the instructor had mentioned the importance of eye contact prior to P1, and this seems to have resulted in everyone commenting on eye contact in the P1 assessment. Students also noted the speech rate, naturalness, nervousness/confidence, voice volume, and non-verbal language.

In contrast to Content and Delivery, less than half of entries addressed Language and Organization, 20 and 17 respectively. In the area of Language, the students mentioned grammar, vocabulary, pronunciation, and fluency, but hardly gave any specific examples. In the Organization category, students commented on introduction, supporting facts, conclusion, transition, or the degree of being organized.

Second, for SA, a similar tendency of focusing more on the Content and Delivery rather than Organization and Language was found as in PA. For Content, six students commented, mentioning interest level, informativeness, or a frank comment about not having been prepared or rehearsed. As in the Delivery section of PA, everyone commented on eye contact, and six students noted being nervous, comfortable, or confident. Everyone commented on Language, with six students noting vocabulary (difficulty level or pronunciation) or grammar. Also in Organization, only three of the seven students made comments, and of the three, one was about transition.

The co-authors of this paper completed TA independently, one as an outside instructor and the other as the course instructor. The TA differed from PA and SA in two main ways. One, the teachers made comments on all four areas of COLD. Two, the teachers identified specific items under Language in addition to providing general comments. The specific comments included appropriateness of sentence endings, honorific and deferent forms, use of particles, grammar points, pronunciation, vocabulary, etc. Taking all the comments from SA, PA, and TA into consideration, assessment criteria of COLD were sub-divided and clarified for use in the subsequent presentations (see Appendix B).

4.2. P2 and P3 Assessment

Table 1 summarizes the descriptive statistics of the students' responses to the post-presentation survey. Paired t-tests were conducted to determine the statistical significance of the changes in the students' responses with the level of significance set at 0.05. Table 2 reveals that not only did the students of this study become more comfortable, but also they felt that they assessed more fairly and responsibly in the third presentation session.

Table 1. Descriptive Statistics of Students' Responses to Post-Presentation Survey

"Felt Comfortable"				*"Assessed Fairly and Responsibly"*			
P2		P3		P2		P3	
M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
3.76	.79	4.24	.63	3.82	.62	4.07	.61

Table 2. Summary of Paired T-tests for the Results of Post-Presentation Survey

	Survey Item	*Paired Difference*		*t*	*df*	*Sig. (2-tailed)*
		Mean	*SD*			
P2-P3	*"Felt Comfortable"*	-.48	.63	-4.03	27	.00*
	"Assessed Fairly and Responsibly"	-.25	.50	-2.62	27	.01*

* significant at p < .05

Tables 3 and 4 summarize the descriptive statistics of SA, PA, and TA before and after the student training session, i.e., for P2 and P3 respectively.

Table 3. Descriptive Statistics of SA, PA, and TA BEFORE Student Training (P2)

	SA	PA	TA	Total
C	*M* = 3.43 *SD* = .51 *Min.* = 2.50 *Max.* = 4.00	*M* = 4.00 *SD* = .25 *Min.* = 3.75 *Max.* = 4.42	*M* = 3.88 *SD* = .39 *Min.* = 3.38 *Max.* = 4.50	*M* = 3.77 *SD* = .46 *Min.* = 2.50 *Max.* = 4.50
O	*M* = 3.61 *SD* = .63 *Min.* = 3.00 *Max.* = 4.50	*M* = 4.08 *SD* = .25 *Min.* = 3.79 *Max.* = 4.50	*M* = 3.65 *SD* = .36 *Min.* = 3.19 *Max.* = 4.13	*M* = 3.78 *SD* = .47 *Min.* = 3.00 *Max.* = 4.50
L	*M* = 3.43 *SD* = .55 *Min.* = 2.50 *Max.* = 4.00	*M* = 4.05 *SD* = .34 *Min.* = 3.48 *Max.* = 4.46	*M* = 3.36 *SD* = .49 *Min.* = 2.81 *Max.* = 4.00	*M* = 3.61 *SD* = .54 *Min.* = 2.50 *Max.* = 4.46
D	*M* = 3.00 *SD* = .63 *Min.* = 2.00 *Max.* = 3.75	*M* = 3.66 *SD* = .41 *Min.* = 3.31 *Max.* = 4.29	*M* = 3.40 *SD* = .51 *Min.* = 2.69 *Max.* = 4.25	*M* = 3.35 *SD* = .57 *Min.* = 2.00 *Max.* = 4.29
Total	*M* = 3.37 *SD* = .60 *Min.* = 2.00 *Max.* = 4.50	*M* = 3.95 *SD* = .35 *Min.* = 3.31 *Max.* = 4.50	*M* = 3.57 *SD* = .47 *Min.* = 2.69 *Max.* = 4.50	*M* = 3.63 *SD* = .53 *Min.* = 2.00 *Max.* = 4.50

Table 4. Descriptive Statistics of SA, PA, and TA AFTER Student Training (P3)

	SA	PA	TA	Total
C	*M* = 3.75 *SD* = .60 *Min.* = 3.00 *Max.* = 4.75	*M* = 4.24 *SD* = .32 *Min.* = 3.75 *Max.* = 4.58	*M* = 4.13 *SD* = .25 *Min.* = 3.81 *Max.* = 4.50	*M* = 4.04 *SD* = .45 *Min.* = 3.00 *Max.* = 4.75
O	*M* = 3.57 *SD* = .84 *Min.* = 2.25 *Max.* = 4.75	*M* = 4.28 *SD* = .28 *Min.* = 3.81 *Max.* = 4.63	*M* = 4.06 *SD* = .48 *Min.* = 3.63 *Max.* = 4.88	*M* = 3.97 *SD* = .63 *Min.* = 2.25 *Max.* = 4.88
L	*M* = 3.82 *SD* = .64 *Min.* = 3.00 *Max.* = 4.75	*M* = 4.35 *SD* = .32 *Min.* = 3.77 *Max.* = 4.69	*M* = 3.47 *SD* = .30 *Min.* = 2.94 *Max.* = 3.81	*M* = 3.88 *SD* = .57 *Min.* = 2.94 *Max.* = 4.75
D	*M* = 3.43 *SD* = .83 *Min.* = 2.25 *Max.* = 4.50	*M* = 3.86 *SD* = .36 *Min.* = 3.44 *Max.* = 4.50	*M* = 3.67 *SD* = .44 *Min.* = 3.06 *Max.* = 4.13	*M* = 3.65 *SD* = .58 *Min.* = 2.25 *Max.* = 4.50
Total	*M* = 3.64 *SD* = .71 *Min.* = 2.25 *Max.* = 4.75	*M* = 4.18 *SD* = .36 *Min.* = 3.44 *Max.* = 4.69	*M* = 3.83 *SD* = .45 *Min.* = 2.94 *Max.* = 4.88	*M* = 3.89 *SD* = .57 *Min.* = 2.25 *Max.* = 4.88

As shown in Tables 3 and 4, the total mean score of P2 was 3.63 and the total mean score of P3 was 3.89. Also, the mean scores for SA, PA, and TA were consistently higher on P3 (SA: M = 3.64 > 3.37; PA: M = 4.18 > 3.95; TA : M = 3.83 > 3.57). Student performance might have improved from the second presentation to the third presentation session. In addition, the students seemed to have improved in all aspects of the evaluation sub-criteria (C : M = 3.77 → 4.04; O : M = 3.78 → 3.97; L : M = 3.61 → 3.88; D : M = 3.35 → 3.65). Furthermore, the mean scores of every cell were higher on P3, with the exception of the mean score of SA on the Organization category (M = 3.61 → 3.57). This is parallel to the interview result where six students answered that the presentations helped them improve oral skills to a certain degree. Not only did the positive comments include language skills (e.g., pronunciation, formal language, honorific usage, and pause), but presentation skills in general such as the preparation process and comfort level. Figures 1 and 2 graphically display changes of the mean scores from P2 to P3 in terms of the three different types of assessment as well as the four different evaluation sub-criteria respectively.

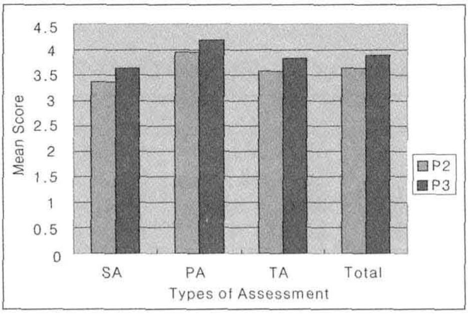

Figure 1. Mean score changes from P2 to P3 (SA, PA, TA, and Total)

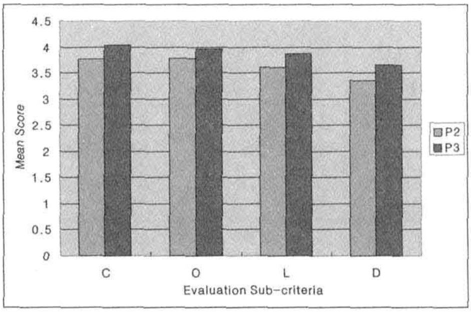

Figure 2. Mean score changes from P2 to P3 (COLD)

Overall, all the research participants tended to give higher marks on the subcategories of Content and Organization, whereas they tended to give lower marks on the subcategories of Language and Delivery. The two categories of Language and Delivery appeared to be more closely related to their ability to use formal language appropriately. Also, in both presentation sessions the total mean scores were lowest in the Delivery category. However, when we look into SA, PA, and TA in more detail, teachers were found to give the lowest scores on the Language component. In contrast, it was the Language component that students scored highest especially in the case of P3. It may be that the students were not ready to evaluate their own as well as their peers' performance in terms of the linguistic appropriateness due to their limited proficiency level. This can be seen in the interview answers when asked if they feel that their ability in identifying strengths and weaknesses in presentations has improved, two students noted that evaluating pronunciation and the correct usage of honorifics was difficult, as well as having had difficulty in evaluating grammar and unfamiliar vocabulary. In addition, two other comments were that they did not feel comfortable evaluating their peer classmates, which seems to have affected the PA scores. Also, when asked about TA, the students commented that "the teacher picks up on things you don't pick up," "TA was helpful," and "I realized I had mispronounced some of the words incorrectly," showing that TA contained points that they were unable to pick up due to their lower proficiency level.

Students tended to underrate their own performance, whereas they tended to overrate their peers' performance. The total mean scores of PA were the highest, followed by those of TA and SA in both presentation sessions. But again, in regard to the Language category, SA scores were higher than TA scores on both the second and third presentation sessions.

In order to test whether the three assessment types were closely associated with each other, the Pearson product-moment correlation coefficients were calculated. Tables 5 and 6 show that in both P2 and P3, overall, SA, PA, and TA are positively correlated with each other. When the subcategories of C, O, L, and D are considered further, the correlation coefficients for the second presentation task ranged from r = -.19 (between SA and TA on the Language category) to r = .70 (between PA and TA on the Language category). In the case of the third presentation task, however, the correlation coefficients ranged from r = .11 (between SA and TA on the Language category) to r = .96 (between PA and TA on the Content category), which suggests that the overall tendency of positive correlation among the three assessment types had strengthened from the second to the third presentation session.

Table 5. Correlation Coefficients for SA, PA, and TA BEFORE Student Training (P2)

	SA-PA	SA-TA	PA-TA
C	.22	.69	.25
O	.47	.63	-.04
L	.17	-.19	.70
D	.56	.69	.63
Total	.49*	.44*	.45*

* significant at p < .05

Table 6. Correlation Coefficients for SA, PA, and TA AFTER Student Training (P3)

	SA-PA	SA-TA	PA-TA
C	.42	.57	.96*
O	.85*	.62	.74
L	.42	.11	.55
D	.63	.75*	.80*
Total	.58*	.42*	.53*

* significant at p < .05

5. Discussion

The students of this study performed a series of oral presentation tasks, whose primary purpose was to help students achieve one of the course objectives of being able to use the Korean language with high formality. It was found that students felt that their presentation skills had improved from the first to the final presentation session. They also reported that they benefited from task performance in terms of their ability to use honorific and formal language expressions, which are important factors for Korean learners in becoming users of advanced proficiency. The descriptive statistics revealed that ratings of their performances had improved in every aspect of assessment sub-criteria: their ability to select interesting and informative topics, to organize presentation content in an appropriate way, to have a command of a language suited for presentations, to make natural delivery without reading from the notes, etc. These results suggest that presentation tasks can be a useful vehicle for learners to develop their language skills.

This study employed self-assessment and peer-assessment as alternative assessment procedures to examine the extent of improvement the participants achieved in their presentation tasks. In order to cope with subjectivity in evaluating behavior, the study encouraged learner involvement in assessment, which was operationalized as students' participation in the development of assessment sub-criteria of oral presentation task performance. The interview results showed that students regarded the evaluation processes as helpful, interesting, and motivating experiences. In addition, the students were provided with three periods of training and discussion opportunities of approximately 55 minutes in total, which resulted in a high comparability among the three assessment types.

Along with the investigation into the validity of SA and PA, some noticeable marking behaviors of students deserve further consideration. First of all, students tended to overrate their peers but underrate themselves as compared with teachers. This tendency appeared to be ascribed to their attitudes to each type of assessment procedures. In line with many earlier reports (e.g., Cheng & Warren, 2005; Falchikov, 1995; Patri, 2002; Williams, 1992), the students of this study felt less comfortable and more uncertain with PA than with SA. Students' comments from in-depth interview about PA experiences included, "I tried not to write negatively since we're all friends, but I needed to write honestly in order to be helpful," "I didn't want to offend anyone," "I didn't make too many comments because I don't know (if) I am in the right position to give such comments." As AlFallay (2004) states, "assessment is a multifaceted process, which is affected by various psychological and personality traits of the raters" (p. 407). Teachers and researchers should be cautious that this over- and/or under-marking behaviors of students may undermine the validity of assessments (Patri, 2002, p. 110).

Second, when we look at the total mean scores of all the research participants, they tended to underrate in the sub-criteria of Language and Delivery, which are closely related to their ability to use formal language. However, closer examination at the student marking behavior on the Language component revealed that students were generous not only to themselves but also to peers with their marks on the Language component. On the contrary, teachers regarded presenters' Language ability as most problematic of the four sub-criteria. This leads us to conclude that for learners with limited language proficiency, the validity of SA and PA might be questionable. They might not be ready in their developmental stages to assess their own and their peers' oral language performance in light of accurate use of grammar, appropriate use of language such as formal endings, honorifics, and fluent use of language with an acceptable pronunciation.

Third, the overall tendency of positive correlation among the three assessment types had strengthened from the second to the third presentation session. This general tendency is also confirmed in the interview results, where five out of seven students responded that their own SA came to match PA and TA in the final presentation session. One student answered that she had expected a lot from herself, was self-conscious, and was hard on herself, while another student stated that when she read PA or TA, she didn't understand initially, but later realized that "it must have been" as stated in the teacher's comment.

The evidence from this study has generated some implications for assessment practices in language pedagogy. Although learners came to be perceived as active organizers who are expected to gain much when they can participate in all activities happening in the classroom (Lee & VanPatten, 2003), much suspicion still remains concerning the idea that acknowledges them as central sources of assessment. Instead, teachers have been commonly assumed as the only reliable and valid sources of assessment. However, this study has led to a possibility that SA and PA can be effectively employed as alternative assessment tools if provided with adequate learner practice and training as well as with students' active involvement in defining assessment criteria. SA and PA may be useful in necessitating a move away from the traditional practices, which clearly have limited scope for developing student responsibility and autonomy, and toward alternatives (Orsmond & Merry, 1996, p. 240).

6. Conclusion

Assessment is a critical component of not only second language learning, but education in general. A call for alternative assessment has brought attention to SA and PA. In this study we have examined the effects of learner involvement in Korean learners' self- and peer-assessment of oral presentation task both quantitatively and qualitatively. With the amount of training and learner involvement conditioned in this study, we may not be able to come to a robust conclusion that SA and PA can be highly comparable with TA. However, as Orsmond and Merry (1996) and Cheng and Warren (2005) suggest, it should be noted that the applicability of SA and PA may not be determined simply by their match or mismatch with TA. If reflecting on the advantages SA and PA may have as alternative assessment procedures, "it is far better to take the risk over the marks than to deprive students of the opportunity of developing the important skills of making objective judgments about the quality of their own work (and that of their peers) and of generally enhancing their learning skills" (Orsmond et al., 1997, p. 358). As Cheng and Warren (2005) put it, SA and PA could work effectively "if the teacher is more concerned with the long-term, cumulative educational benefits rather than simply the immediate success or failure of students' attempts to imitate or supplement the assessment behavior of their teacher" (p. 112). In order to have a more complete view of the benefits of SA and PA, longitudinal studies along similar lines can help improve these preliminary but promising findings.

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, 407-425.

Angelo, T. A. and P. K. Cross. (1993). *Classroom Assessment Techniques: A Handbook for College Teachers.* San Francisco: Jossey-Bass.

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

Boud, D. (1989). The role of self-assessment in student grading. *Assessment and Evaluation in Higher Education* 14, 20-30.

Cheng, W. and M. Warren. (2005). Peer assessment of language proficiency. *Language Testing* 22, 93-121.

Eisner, E. W. (1999). The uses and limits of performance assessment. *Phi Delta Kappan* 80, 658-660.

Falchikov, N. (1986). Product comparison and process benefits of collaborative peer group and self-assessment. *Assessment and Evaluation in Higher Education* 11, 146-166.

Falchikov, N. (1995). Peer feedback marking: Development 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, 289-300.

Haertel, E. H. (1999). Performance assessment and education reform. *Phi Delta Kappan* 80, 662-673.

Heilenman, L. (1991). Self-assessment and placement: A review of the issues. In R. Teschner, ed., *Assessing Foreign Language Proficiency of Undergraduates* (pp. 93-114). Boston: Heinle & Heinle.

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

Kwan, K. and R. Leung. (1996). Tutor versus peer group assessment of student performance in a stimulation training exercise. *Assessment and Evaluation in Higher Education* 21, 239-249.

LeBlanc, R. and G. Painchaud. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly* 19, 673-687.

Lee, J. and B. VanPatten. (2003). *Making Communicative Language Teaching Happen.* McGraw-Hill.

Long, M. H. and J. M. Norris. (2000). Task-based language teaching and assessment. In M. Byram, ed., Routledge *Encyclopedia of Language Teaching and Learning* (pp. 597-603). New York: Routledge.

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

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice* 22, 13-25.

Norris, J. M. (2000). Purposeful language assessment: Selecting the right alternative test. *English Teaching Forum* 38, 18-23.

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing* 19, 337-346.

Norris, J. M., Brown, J. D., Hudson, T. D., and W. Bonk. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing* 19, 395-418.

Oldfield, K. and M. Macalpine. (1995). Peer and self-assessment at tertiary level: An experimental report. *Assessment and Evaluation in Higher Education* 20, 125-132.

Orsmond, P. and S. Merry. (1996). The importance of marking criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education* 21, 239-251.

Orsmond, P., Merry, S., and K. Reiling. (1997). A study in self assessment: Tutor and students' perceptions of performance criteria. *Assessment and Evaluation in Higher Education* 22, 357-367.

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

Sohn, H. (2001). *The Korean Language.* Cambridge: Cambridge University Press.

Sohn, H. and E.-J. Lee. (2003). *Integrated Korean, Advanced Intermediate.* Honolulu: University of Hawai'i Press.

Stanley, K. (2003). A question of definitions: An investigation through the definitions and practices of communicative and task-based approaches. *TESL-EJ* 7(3).

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

Stefani, L. (1998). Assessment in partnership with learners. *Assessment and Evaluation in Higher Education* 23, 339-350.

Taras, M. (2001). The use of tutor feedback and student self-assessment in summative assessment tasks: Towards transparency for students and for tutors. *Assessment and Evaluation in Higher Education* 26, 605-614.

Taras, M. (2002). Using assessment for learning and learning from assessment. *Assessment and Evaluation in Higher Education* 27, 501-510.

Wiggins, G. P. (1993a). *Assessing Student Performance: Exploring the Purpose and Limits of Testing.* San Francisco: Jossey-Bass Publishers.

Wiggins, G. P. (1993b). Assessment: Authenticity, context, and validity. *Phi Delta Kappan* 75, 200-214.

Williams, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education* 17, 45-58.

APPENDIX A
OPEN-ENDED ASSESSMENT FORM (P1 / PA)

Your Student # ____________

Presenter's Student # _______________

Topic of the Presentation ____

Date ___________________________________

In a paragraph or two, please give your assessment of the presentation.

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

APPENDIX B

ASSESSMENT FORM (P2 & P3 / PA)

Your Student # ____________

Presenter's Student # ____________

Use the following scale when assessing your fellow students:

Assessment Criteria	Assessment Scale	More Comments
I. Content & Preparation (Overall Score: / 5)		
1. Interest and relevance value of topic including presenter's opinion	↔ 1 poor 2 below average 3 average 4 above average 5 excellent	
2. Informativeness and sufficient quantity of content	↔ 1 poor 2 below average 3 average 4 above average 5 excellent	
3. Preparedness (evidence of rehearsal)	↔ 1 poor 2 below average 3 average 4 above average 5 excellent	
4. Coherence and clarity	↔ 1 poor 2 below average 3 average 4 above average 5 excellent	

II. Organization (Overall Score: / 5)		
1. Introduction (mentioning of topic and overview)	⟷ 1 poor 2 below average 3 average 4 above average 5 excellent	
2. Main body (supporting details/examples, clarity)	⟷ 1 poor 2 below average 3 average 4 above average 5 excellent	
3. Conclusion (brief summary of the presentation)	⟷ 1 poor 2 below average 3 average 4 above average 5 excellent	
4. Smoothness of transition	⟷ 1 poor 2 below average 3 average 4 above average 5 excellent	

III. Language use of words and expressions learned in class (Overall Score: / 5)		
1. Accuracy (accurate use of grammar)	⟷ 1 poor 2 below average 3 average 4 above average 5 excellent	
2. Appropriateness (appropriate use of language such of formal endings, honorifics, and use of words and expressions learned in class)	⟷ 1 poor 2 below average 3 average 4 above average 5 excellent	

3. Fluency (pauses in appropriate places, flow)	1 poor 2 below average 3 average 4 above average 5 excellent	
4. Pronunciation	1 poor 2 below average 3 average 4 above average 5 excellent	

IV. Delivery (Overall Score: / 5)		
1. Naturalness of delivery (not read or fully memorized, appropriate speech rate)	1 poor 2 below average 3 average 4 above average 5 excellent	
2. Confidence (not being overly dependent on notes)	1 poor 2 below average 3 average 4 above average 5 excellent	
3. Rapport with and sensitivity to audience	1 poor 2 below average 3 average 4 above average 5 excellent	
4. Non-verbal interaction with the audience (eye contact, facial expressions, gestures, not too many unnecessary filler words)	1 poor 2 below average 3 average 4 above average 5 excellent	
General Comments		

APPENDIX C
POST-PRESENTATION SURVEY

We've used a scale when assessing your peer students for the second presentation. Using the chart below, please carefully rate how you felt about peer assessment. Use the scale of 1 to 5, with 1 being strongly disagree, 2 disagree, 3 neutral, 4 agree, and 5 strongly agree. Circle the most appropriate one.

Assessment Criteria	I felt comfortable in SA and PA on each of the assessment criteiria.	I think I assessed fairly and responsibly in SA and PA.
I. Content & Organization		
1. Interest and relevance value of topic including presenter's opinion	1 2 3 4 5	1 2 3 4 5
2. Informativeness and sufficient quantity of content	1 2 3 4 5	1 2 3 4 5
3. Preparedness (evidence of rehearsal)	1 2 3 4 5	1 2 3 4 5
4. Coherence and clarity	1 2 3 4 5	1 2 3 4 5
II. Organization		
1. Introduction (mentioning of topic and overview)	1 2 3 4 5	1 2 3 4 5
2. Main body (supporting details/examples, clarity)	1 2 3 4 5	1 2 3 4 5
3. Conclusion (brief summary of the presentation)	1 2 3 4 5	1 2 3 4 5
4. Smoothness of transitions	1 2 3 4 5	1 2 3 4 5
III. Language: Use of words and expressions learned in class		
1. Accuracy (accurate use of grammar)	1 2 3 4 5	1 2 3 4 5
2. Appropriateness (appropriate use of language such as formal endings, honorifics, and use of words and expressions learned in class)	1 2 3 4 5	1 2 3 4 5
3. Fluency (pauses in appropriate places, flow)	1 2 3 4 5	1 2 3 4 5
4. Pronunciation	1 2 3 4 5	1 2 3 4 5

IV. Delivery		
1. Naturalness of delivery (not read or fully memorized, appropriate speech rate)	1 2 3 4 5	1 2 3 4 5
2. Confidence (not being overly dependent on notes)	1 2 3 4 5	1 2 3 4 5
3. Rapport with and sensitivity to audience	1 2 3 4 5	1 2 3 4 5
4. Non-verbal interaction with the audience (eye contact, facial expressions, gestures, not too many unnecessary filler words)	1 2 3 4 5	1 2 3 4 5

Sang-Ki Lee
Department of Second Language Studies
University of Hawai'i at Mānoa
1890 East-West Rd.
Honolulu, HI 96822
E-mail: sangki@hawaii.edu

Sumi Chang
Department of East Asian Languages and Literatures
University of Hawai'i at Mānoa
1890 East-West Rd.
Honolulu, HI 96822
E-mail: changhan@hawaii.edu

Received: Jun. 2, 2005
Revised version received: Aug. 15, 2005
Accepted: Sep. 8, 2005