Manuscript published in **M. Nikolov (Ed.),** ***Assessing Young Learners of English: Global and Local Perspectives*** **(pp. 317-338). New York: Springer.**

The Relationships between Peer- and Self-Assessment and Teacher Assessment of Young EFL Learners' Oral Presentations

Yu-Ju Hung, Beth Lewis Samuelson and Shu-Cheng Chen

As the traditional grammar translation approach is being gradually replaced by communicative or task-based approaches, paper-and-pencil tests, commonly used in English classes in Taiwan, do not meet the course goals. Alternative assessment, known for increasing learners' cognitive and meta-cognitive development as well as empowering students to take ownership of their learning, has been practiced extensively in L1 higher education, but neglected in L2 elementary schools. Thus, the purpose of this study is to investigate how peer and self-assessment can be implemented to evaluate young EFL learners' oral presentation and how the students perceive this experience. The study was conducted in two sixth grade classes at a public elementary school in southern Taiwan. After attending a professional development workshop held by the government, a local English teacher practiced peer and self-assessment in her class so as to engage every student in class activities and also to provide an opportunity for them to reflect upon their performance. In the process, the students formed groups of six to discuss and give grades after each individual student's oral report. Three types of data sources were analyzed. The first was the evaluation rubrics from peer groups, each presenting students, and the teacher. Then, a survey, containing 16 closed-format questions and one open-ended question, was administered to elicit the students' perceptions of the assessment process. Also, an interview was done with the teacher. The results show that peer and teacher assessment had strong positive correlation, whereas self- and teacher assessment were moderately correlated. The strength of correlation also varied for each evaluation criterion. Though learners responded positively to the assessing experiences in the questionnaires, they expressed concern that some grades assigned by peers were not fair and a few group members dominated the grading process. The findings shed light on benefits of combining peer and self- assessment and suggest training should emphasize self-assessment, evaluation criteria related to content of the presentation, and students' social skills to work harmoniously in groups. Most of all, students' traditional way of learning should not be neglected.

1 Introduction

As the Ministry of Education in Taiwan has listed communication as one of the main objectives of English instruction in elementary school and encouraged alternative assessment (Ye, 2001), learner-centered instruction has started to gain popularity in EFL classrooms. Peer and self- assessment (hereafter PA and SA) are two forms of classroom assessment that involve students' participation to a great extent. PA is "an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners" (Topping, 2010, p. 62). In PA, students judge the work of their peers whereas students judge their own work in SA (Falchikov & Goldfinch, 2000). PA and SA have been found to motivate students and improve their learning (Dochy, Segers, & Sluijsmans, 1999; Hung, Chen, & Samuelson, under review).

PA and SA can be reciprocal. Students' experiences of critiquing and evaluating in PA informs their SA (Topping & Ehly, 2001). On the other hand, SA unavoidably refers to viewpoints and judgments of others (Boud, 1995). Also, combination of PA and SA has been suggested to prevent over-marking in rating peers and under-marking in rating students' own work (Dochy et al., 1999) though the issue of accuracy still remains questionable. We argue that a combination of PA and SA increases agreement between student and teacher assessment and benefits students' learning.

However, few classroom assessment studies that incorporate both PA and SA have been conducted in EFL contexts, particularly for young learners' oral presentation. Oral SA is more difficult to practice and Harris (1997) suggested it be supplemented by PA. Therefore, the purpose of this study, grounded on observational learning in social learning theory (Bandura, 1971), is to investigate how PA and SA can be implemented to evaluate young EFL learners' oral presentation and how students perceive this assessment experience. The two research questions are

1. What are the relationships like between peer, self-, and teacher assessment?
2. How do students and the teacher perceive the assessment experience?

2 Observational Learning in Social Learning Theory

This study is situated within the framework of observational learning in social learning theory (Bandura, 1971), later reconceptualized as social cognitive theory (Bandura, 1991). In this framework, human behavior is neither driven by inner forces, nor is human behavior shaped by trial and error, as proposed in the conditioning view. -. Rather, the causes of behavior are cognitively mediated by means of a continuous reciprocal interaction between behavior and environmental forces. New patterns of behavior are the causal consequences arising from cognitively mediating the influences of stimuli of given activities. Among the stimulus determinants, learning first occurs through direct experience or by observing the behavior of others. Thus, providing an appropriate model of the target learning behavior is indispensable in the process.

2.1 Learning through modeling

Social learning theory does not accept that learners simply imitate a model's actions, but that they form new response patterns by organizing behavioral elements they observed. This modeling learning is governed by four processes. The first is attentional processes. Learners select from the model's numerous characteristics and attend to the most relevant ones. Associational preferences are another essential factor. Learners associate with members in their social groups. In other words, learners relate to their peers in classroom settings. The second is retention processes. Verbal coding of the observed information facilitates cognitive processing and storage. Also, rehearsals, or actually performing or mentally rehearsing, enhance long-term retention. The third component of modeling involves motoric reproduction processes. Learners first acquire symbolic representations of modeled activities; thus, they achieve approximations of the desired behavior. They refine the new patterns of behavior through self-corrective adjustments according to feedback from their own performance. The fourth factor is reinforcement and motivational processes. Positive feedback or incentives activate the acquired skills to actual performance. Anticipation of positive consequences is one of the best motivators to reinforce and generate an effective, high level of observational learning (Bandura, 1971).

Similarly, students rated their peers' performances based on the criteria in the evaluation rubrics in the present study, so they selectively attended to features of their peers' oral presentations. After each presentation, they discussed and decided the scores on individual assessment criteria as a group. Each group and the teacher then gave oral feedback about the strengths and weaknesses of the presentation. This verbalizing process helped them understand and retain the criteria. The assessing experiences also provided students opportunities for self-reflection by casting themselves in a similar context, a form of mental rehearsal to facilitate their future performance. Afterwards, their SA reinforced their assessment ability for their own presentation and benefited their learning. Self-observation and self-judgment in the process of SA informed leaners how well they were progressing toward their goals and motivated behavioral change (Schunk, 2001).

2.2 Functions of reinforcement

Within the framework of social learning theory, an effective, high-level of observational learning of modeled behaviors is shaped and activated by three functions: informative function, motivational function, and cognitive function.

Informative function of reinforcement indicates learners observe modeled behaviors and conceive what they must do to obtain beneficial consequence. When doing ratings, students reflect by thinking, comparing, contrasting what they observe (Topping, 1998). For motivational function, anticipated consequence and affective factors, such as being empowered to do ratings, serve as best incentives. Cognitively mediated reinforcement offers students opportunities to selectively pay attention to what to reward and ignore. Using evaluation criteria and peer group discussion of the criteria reinforce students' understanding of standards of high quality presentations.

3 PA and SA in L1 and L2 Contexts

Relevant studies of PA and SA have been carried out extensively in various fields in L1 higher education contexts, but fewer studies combine both forms of assessment of target oral performance in L2 contexts, especially with young learners. This section reviews PA or SA in higher education first and then narrows down the scope to discuss empirical studies incorporating both forms of student-assessment with young learners.

3.1 Reviews of PA and SA

The PA process, in which students benefit from social interaction between assessors and assesses, enhances development of cognition and meta-cognition, affect, and social skills (Topping, 1998). Reviews of PA studies find general agreement between student and teacher ratings. Falchikov and Goldfinch (2000) analyzed 48 quantitative studies in L1 settings from 1959 to 1999 and found the mean value of correlation coefficients was 0.69, indicating general agreement between peer and teacher ratings. Consistent with the previous findings, van Zundert, Sluijsmans, and van Merriënboer (2010) reviewed 26 studies of L1 PA from 1990 to 2007 and further pointed out that peer feedback helped students revise their work, higher achievers were more skillful in PA than lower achievers, and students had mixed attitudes toward PA. The problems of friendship marking (Pond, UI-Hag, & Wade, 1995), or addressed as "reciprocity effects" (Panadero, Romero, & Strijbos, 2013, p. 195), and insufficient differentiation (Murphy & Cleveland, 1995), which indicated that learners gave ratings higher than their peers deserved and they tended to give their peers a narrower range of ratings to avoid inaccurate evaluations, were commonly shown in adult learners.

Given opportunities to assess and reflect on their individual progress in SA, learners focus on their own learning, locate their strengths and weaknesses, and take responsibility for their own learning (Harris, 1997).

The review of SA research shows self-appraisal improves students' achievement, though the results for self- and teacher agreement are not as good as for PA (Blanche & Merino, 1989; Ross, 2006). SA of oral skills is found to be more difficult because speaking can be highly intangible (Harris, 1997). Self-ratings may be affected by subjective errors due to past academic record, peer or parental expectations (Blanche & Merino, 1989). Cultural factors, such as the pressure to display overt modesty, which is valued in Chinese culture, may make students more critical of their own performance (Chen, 2008; Oscarson, 1997). In contrast, Iranian students are lenient when rating themselves since overt or false modesty concerning one's accomplishments is not accentuated in their culture (Esfandiari & Myford, 2013). Young children tend to over-estimate due to their wishful thinking and lack of the cognitive skills to evaluate their abilities accurately (Ross, 2006).

The above reviews show benefits as well as potential problems of PA and SA. Dochy et al. (1999) argued that incorporating both types of student assessment could overcome the defects of over-marking and under-marking. However, the following studies show that this proposal still remains in question and that empirical studies are needed to verify this argument.

3.2 Combination of PA and SA

In studies that combine PA and SA of oral performance in L1 universities, student and teacher ratings show disagreement (De Grez, Valcke, & Roozen, 2012; Fallows & Chandramohan, 2001; Langan, Shuker, Cullen, Penney, Preziosi, & Weater, 2008) and agreement (Lanning, Brickhouse, Gunsolley, Ranson, & Willett, 2011). The disagreement between student and teacher ratings might be due to different interpretations of evaluation criteria between them (De Grez et al., 2012). Particularly in Asian contexts, low achievers over-marked, and high achievers under-marked. Students' hesitation or lack of confidence to distinguish their peers' performances resulted in a narrower range of rating their peers. Students also reported they could not pay full attention to their peers' performance because they needed to do peer-marking while watching the performance (Langan et al., 2008). This result contradicts the findings of previous studies and calls into question the idea of learning from modeling because students are so focused on assessing their peers that they may not be able to observe the performance attentively.

The positive effects of proper training, involving students in constructing evaluation criteria, providing more opportunities for student assessments, and combining PA and SA with teacher feedback are shown in other empirical studies in L1 and L2 contexts though the tasks are not on oral performance. In Orsmond, Merry, and Reiling's (2002) study, using exemplars to discuss criteria helped students understand what was expected and could create agreement between students and teachers though better agreement was observed between PA and teacher assessment than SA and teacher assessment. Students appeared to be more objective and look at product, the presentation itself, in rating peers, but more subjective and look at the process, how they prepared for the presentation, in rating themselves. Nevertheless, though benefits of student assessments were recognized, they could not replace teacher assessment. The appropriate combination of PA, SA, and teacher assessment had the best impact on student learning of assessment as well as target skills (Birjandi & Tamjid, 2012; Murakami, Valvona, & Broudy, 2012).

3.3 PA and SA with young learners

Student assessment has been found to have a positive effect on young learners' achievement, but an age-related difference appears to be a factor. Ross, Hogaboam-Gray, and Rolheiser (2002) found that 5th and 6th graders who received self-evaluation training had a higher math achievement than who did not. Butler (1990) compared ratings of children at ages 5, 7, and 10 with adult judges after they copied drawings. Young learners were interested and capable of comparing drawings with standards. However, when learners were put in competitive condition, the desire to outperform others and difficulties in evaluating relative abilities caused inflated perceptions of their own work and decreased their interest.

Butler and Lee (2006) compared 4th and 6th graders' SA with teacher assessment and results of standardized tests in Korea. The study showed that the 6th graders out-performed 4th graders in terms of student assessment accuracy. The researchers concluded that cognitive development influenced young learners' self-appraisal ability. In another study of 6th graders in English class in Korea, repeated SA improved students' assessing ability as well as English performance on objective tests (Butler & Lee, 2010). On the other hand, the agreement between SA and teacher assessment in the control group decreased over time. A possible interpretation of this decline was that young learners were more positive on their academic performance, but the perception decreased by the time they finished elementary school.

In Mok's (2010) study, four secondary students expressed serious concerns that they were not good enough to evaluate their peers, even though they agreed PA helped them reflect upon their own performances. Mok called for preparation of the students both methodologically and psychologically for the role of peer assessor. Hung, Chen, and Samuelson (under review) examined group PA of 4th to 6th graders' oral performance in EFL classes in Taiwan. The results showed that the 5th and 6th graders were able to assess their peers as their teacher did, whereas the 4th graders were not. The majority of the students in all levels reported they enjoyed playing the role of assessor and indicated this process benefited their subsequent performance and English learning. However, challenges of accepting diverse opinions and conducting discussions of evaluating their peers within groups, particularly for the 4th graders, were indicated.

Though there are some preliminary findings of practicing PA and SA with young learners in the related literature, the effect of combining the two remains uninvestigated and therefore is the main focus of this empirical study.

4 Research Method

This classroom-based research used both quantitative and qualitative data to reveal the assessment process as well as the opinions of the students and their teacher. Author 3 worked collaboratively with two university researchers to plan and implement student assessment procedures in her class. Author 1 observed all classes in which student assessment was conducted. Author 2 assisted with research data analysis, and her prior experience as an English teacher in southern Taiwan helped her to be familiar with the educational context of the study.

4.1 Setting and participants

The setting for this study was a public elementary school in southern Taiwan. The school was established in 1996 to serve a new high socioeconomic status (SES) suburban community. The total student population was about 800 students, divided into 30 classes (grades 1-6). This school was regarded as a high performing school where the teachers as well as the students had received awards for excellence from the local government and the Ministry of Education.

Approximately 90% of the students were Taiwanese; 10% were Hakka (an ethnic Chinese group comprising 15-20% of Taiwan's population); or immigrants from provinces in Mainland China or other countries. When the study was conducted, students were required to study English from 3rd grade in elementary school (age 9) based on the national policy. However, local educational policy promoting English proficiency required all students at this school to start English courses from the second grade (age 8).

4.2 The teacher and the students

Author 3 held a MA degree of English teaching and had been teaching English at elementary school for 14 years. After attending a workshop of student assessment held by the Ministry of Education, she carried out the PA and SA activities in two six-grade classes. These classes were selected because they were taught by the same teacher. Sixty-nine students participated in the study, with three students excluded due to absences. Forty-two were female students and twenty-seven were male. All of the students began learning English in the second grade and had received two 40-minute English classes every week. In addition to the formal English instruction in elementary school, 58% of the students (N=40) started to learn English from tutors or in private institutes before entering elementary school, and an additional 16% of them (N=11) started in 1st grade. Approximately 96% of the participants (N=66) learned English out of class when this study was conducted.

Based on routine placement tests in the beginning of the semester and the students' final English grades the previous semester, all 6th graders, had been divided into advanced, intermediate, and basic levels. The participants in the current study were assigned to advanced classes. For the purpose of the study, the students were arranged in groups of six for PA. There were twelve peer groups in the two classes, six groups in each class.

4.3 The classroom atmosphere

Author 3 emphasized communicative competence through simple daily conversations. Grammar was not focused on. The students were required to take an oral exam and a written exam to fulfill the course requirement. The instructional approach involved a lot of teacher-student and student-student interaction, role-play, and English games. Because the majority of the students had also been taught by Author 3 in 5th grade, they were quite accustomed to these activities and felt comfortable to talk and participated to a great extent in English class.

4.4 PA and SA procedures

Training students to ensure they are aware of the objectives and procedures of the assessment and understand evaluation criteria is the key to successful PA and SA activities. Several important steps mentioned in the literature include clarifying the purpose of the kind of assessment done and expectations of the students as assessors; involving participants in developing assessment criteria; providing practice and examples of student performance; providing written checklists or guidelines, specifying activities and timescale; giving feedback; and examining the quality of feedback (Oscarson, 1997; Topping, 2009). Accordingly, the researchers designed the following training procedure. The entire procedure lasted seven weeks to complete for each class: two class periods per week and 40 minutes per class period. After Author 3 taught the textbook content in each class, she spared approximately one third of the course time for the student assessment activity. Training took one class period. The process writing activity took three weeks. Presentations took three weeks. Six to eight presentations were done per class.

Step 1. Introducing PA and SA

Author 3 informed students that PA and SA would be used to evaluate their oral presentations. Students' final grades would include peer, self- and teacher ratings. The purpose and rationale of student assessment were introduced. Students were told that evaluation should be decided from different perspectives, not only by their teacher, but also by their fellow students. When they did PA, they were learning English from others at the same time. They could reflect on their own performance by rating others and themselves and improve their own future presentation. Author 3 encouraged the students to take responsibility for the process and learn from the assessing process. After Author 3 introduced PA and SA, students moved on to prepare for their oral presentations.

Step 2. Preparing oral presentations

This class used the English textbook, *Enjoy 10*, issued by the local Bureau of Education (Shen et al., 2001). The first unit covered the topic of traveling, and the students had just finished their summer vacation. Author 3 decided to use "My Summer Vacation" as the presentation topic. Since the English level of this group of students was still at the beginner's stage, Author 3 guided the students to draft their presentation content via process writing. After the students composed draft 1 at home and submitted it to Author 3, she indicated the parts that the students could elaborate and taught them how to look up English words online. In the second draft, Author 3 underlined obvious language errors. In the final draft, Author 3 corrected language errors that the students could not revise by themselves. Figure 1 was a final draft by one of the students. In the presentation, the student memorized the content and recited it in front of their classmates.

> My Summer Vacation
>
> In my summer vacation, I went to day care center every day. On Saturday in July, the day care center took me and many other students out. We did some interesting things. We saw a movie, Despicable Me 2, played bowling, and then ate dinner. I enjoyed that movie. I had fun playing bowling. The dinner was great. I was very happy in my summer vacation.

Figure 1 Student Writing Sample

Step 3. Discussing evaluation criteria

Involving students in the development of evaluation criteria has been recommended in the literature to help learners understand what constitutes a good presentation to develop a sense of ownership (Harris, 1997; Topping, 2009). Author 3 discussed the evaluation criteria with the whole class and decided on them together (see Figure 2). The students agreed that the four criteria should be weighted differently. From Author 3's previous experience of practicing student assessment, students tended to focus on their peers' weaknesses instead of strengths, so strengths and suggestions were used in the comment to lead the students to pay more attention to their peers' strengths and give feedback constructively. Finally, Author 3 discussed with the students what should be considered the standard for each criterion.

Evaluation Rubric	
Voice (6 points)	
Content (6 points)	
Interaction with audience (6 points)	
Body language & facial expression (2 points)	
total (20 points)	
Strength:	
Suggestion:	

Figure 2 Evaluation Criteria

Step 4. Presenting and evaluating

Right before the first presentations, the students reviewed the evaluation criteria. After each presentation, the audience discussed their classmate's performance within their groups and assessed their peer by deciding the grades as a group. Meanwhile, each presenting student did a SA using the same rubric. Then the teacher and each student group gave oral feedback on the performance. The assessment of all presentations followed the same pattern. Since the students' English abilities were developing, the discussion within groups and in the whole class was conducted in Chinese.

Step 5. Reflecting

Author 3 calculated the final scores across groups and compiled all the comments from each group. In the next class, she gave each group its results. She then led the whole class in a reflective discussion on the assessment process.

4.5 Data sources

In addition to peer, self-, and teacher ratings for each presentation, data included a post-assessment survey filled out by the students and a teacher interview. The survey items and their Chinese translations were examined by Author 3 to establish the content validity, drawn on a subject matter expert's judgment of whether a measure includes the appropriate content for the construct it aims to measure (Cohen, & Swerdlik, 2005). Chinese versions of the questionnaire along with a parental consent form were given to the students. Only students who completed both the survey and returned the consent form were included (N=69). The design of the five-point Likert scale questionnaire to examine the ratings and interactions between assessors and assesses as well as among team members was framed by social learning theory (Bandura, 1971). In addition to students' demographic information, the items were constructed on the basis of three functions of reinforcement in observational learning, including informational function (Items 1-7), motivational function (Items 8-11), and cognitive function (Items 12-16). One open-ended question elicited the students' general reflection on this process (see Table 1).

Table 1 Student survey

information function
1. I paid more attention to my classmates' presentations when I evaluated them.
2. I learned English from evaluating my classmates' presentations.
3. I learned how to do a good oral presentation from rating my classmates.
4. My classmates' feedback was helpful to my presentation.
5. I could reflect my own presentation and think how to improve from evaluating myself.
6. I learned how to give clear concrete suggestions from giving my classmates feedback.
7. I learned how to encourage the presenter from giving my classmates feedback.
motivational function
8. I liked this assessing activity.
9. I could assess my classmates objectively.
10. I could assess myself objectively.
11. My classmates could assess me objectively.
cognitive function
12. The whole class discussion of evaluation criteria helped me understand how to prepare my oral presentation.
13. Each member had chance to express their own opinions in group discussions.
14. My group members accepted each other's opinions in group discussions.
15. My opinions had been accepted in group discussions.

16. I had accepted my group members’ opinion in group discussion.

The semi-structured teacher interview probed the teacher’s perceptions of this assessment practice. The questions included the benefits and difficulties she encountered and how she would expect it to be modified in future classes. The interview was recorded and transcribed.

4.6 Data analysis

4.6.1 Rubric data

A Paired Samples T-Test was used to compare differences between mean scores of peer, self-, and teacher ratings to reveal whether students’ perception of their performance accorded with their teacher (Isaac & Michael, 1995). Correlation was used to analyze agreement of total scores and scores of each evaluation criterion between peer, self-, and teacher ratings. Agreement was confirmed if the peer or self-ratings lay within one standard deviation of the teacher’s ratings (Kwan & Leung, 1996). The maximum and minimum scores of PA, SA and teacher assessment were also compared to examine the range of their ratings.

4.6.2 Questionnaires

Descriptive analysis was used to tabulate numbers, percentages, and mean scores of the results of the questionnaires. Cronbach's alpha coefficient for the 16 items is .873, suggesting high reliability of the questionnaire.

4.6.3 Open-ended questions

Students’ responses to the open-ended question in the survey and the teacher interview were coded using three functions of reinforcement of observational learning. Author 1 and Author 3 coded all the data independently. A Kappa measure of the two raters’ coding was greater than 0.85, indicating acceptable inter-rater reliability (Landis & Koch, 1977). Agreement on each coding was reached through discussion.

5 Results

We will present our findings in terms of each of the research questions given at the beginning of this article. The peer, self-, and teacher ratings are used to show the correlations between their evaluations, and the student survey and teacher interview are used to delineate their perceptions.

5.1 Agreement of PA, SA, and teacher assessment

The analyses of PA, SA, and teacher assessment reveal peer, self-, and teacher ratings were correlated to a certain extent in the present study. Over-marking, under-marking, and range restriction, which appeared in previous studies of PA or SA, did not exist in this study. As Table 2 shows, the ranges of peer and self- ratings are 9-20 and 7-20, respectively; whereas the range of teacher rating is 12-20. The ranges of both peer and self- ratings are larger than teacher ratings. The mean differences between peer and teacher ratings and between self- and teacher ratings lay within one standard deviation of the teacher ratings, which indicates agreement between peer and teacher ratings as well as self- and teacher ratings (Kwan & Leung, 1996). Though the mean scores of peer- and self- ratings are slightly lower than the mean score of the teacher ratings, Paired sample T tests reveal no significant differences between peer and teacher ratings (p > .05) and between self- and teacher ratings (p > .05). As displayed in Table 3, the Pearson correlation coefficient between peer and teacher ratings is .73 (p < .01), while the correlation coefficient between self- and teacher ratings is .48 (p < .01). A correlation of 0.5 is large, 0.3 is moderate, and 0.1 is small (Cohen, 1988). The results show that PA and teacher assessment had a strong positive correlation, whereas the correlation between SA and teacher assessment was moderate and positive. Both correlations were significant.

Table 2 Descriptive statistics for peer, self-, and teacher ratings

	Mean	Std. Deviation	Minimum	Maximum	N
Peer Rating	16.51	1.61	9	20	69
Self-Rating	16.09	2.51	7	20	69
Teacher Rating	16.66	2.15	12	20	69

Table 3 Correlation between peer, self-, and teacher ratings

Total			
		Peer	Self
Teacher	Pearson Correlation	.73**	.48**
	Sig. (2-tailed)	.00	.00
	N	69	69
Voice			
		Peer	Self
Teacher	Pearson Correlation	.76**	.49**
	Sig. (2-tailed)	.00	.00
	N	69	69
Content			
		Peer	Self
Teacher	Pearson Correlation	.44**	.34**
	Sig. (2-tailed)	.00	.00
	N	69	69
Interaction with audience			
		Peer	Self

Teacher	Pearson Correlation	.60**	.28*
	Sig. (2-tailed)	.00	.02
	N	69	69
Body language and facial expression			
		Peer	Self
Teacher	Pearson Correlation	.40**	.25*
	Sig. (2-tailed)	.00	.04
	N	69	69

In the interview, the teacher also stated her observation of the difference between PA and SA. Some students might have over-marked themselves because they subjectively took into account their effort. Author 3 thought that the students' SA of their effort was a good supplementation to other assessments, since it was difficult to tell the students' preparation process. As she stated in the interview,

> When a student rated their peers' performance, he watched the performance of the student critically. When the presenter evaluated himself, he must have thought 'How much effort do I put into this? How is my performance in my point of view?' He evaluated his own performance from his own perspective, not from the perspective of an outsider. I could compare the differences of the evaluations from two perspectives. (Teacher Interview)
> 當學生幫學生打分數時，他們站的是比較批判的角度去看這個學生的表現，可是如果是站在這個 presenter 的角度，我從中努力了多少，我看到我自己的表現是多少，那我想要看中間的不同點，而不是只是站在 outsider 的角度，要從我自己的角度去看。

Unlike previous studies that indicated students tended to under-mark themselves because modesty was valued in Chinese culture (Chen, 2008), only a small number of students under-marked themselves in this study, and that was partially because they set high standards for themselves. The teacher also commented on this phenomenon in her interview,

> When judging oneself, one always knew all the hard work done and this could prompt a student to rate him/herself more generously. Few students marked themselves really low. These were special cases. They gave themselves really low grades, but their performances were very good according to the teacher's scores. This might be because they had high expectations of themselves. It might also reflect their desire to display humility about their accomplishments. But these were the minority. From

> what I observed, most of the students did not rate themselves very differently from their peers' evaluations of them. (Teacher Interview)
> 有些學生會把自己的分數打得比較高，中間的差異性在於他知道自己付出多少，有些學生把自己的分數打得特別低，這幾個特別 case，打得特別低的，但表現得很好的，就是自我要求很高，就是他很習慣性的 too humble，但是這種情況比較少，大部分打出來的，觀察之下，其實跟同儕打的差距性也沒有很大。

Table 3 also shows correlations between peer, self-, and teacher ratings for each evaluation criterion. Though all of the criteria are positively correlated between peer, self-, and teacher correlation, slight differences exist in correlation between PA and teacher assessment. For the criteria of voice and interaction with audience, PA and teacher assessment are strongly correlated (r = .76 and r = .60); in contrast, the correlations of content and body language and facial expression are relatively weak (r = .44 and r = .40). The criteria of voice and interaction with audience are probably easier to observe and evaluate. For the content, the students might not have comprehended their peers' presentation completely or they might have had different standards from the teacher. The total number of points for the criterion of body language and facial expression was only 2, which help to explain the weak correlation.

5.2 Reinforcement functions of PA and SA

5.2.1 Informative function

The students recognized what they had learned from the assessing activity. Approximately 95% of the students strongly agreed or somewhat agreed that they paid attention to their peers' presentation, learned English, learned how to do a presentation, and gave and got feedback to improve themselves (see Table 4). As one student stated in the survey,

> This was a great activity! By rating our classmates' presentations, we gave ratings, and we also learned to accept others' opinions. When others evaluated us, they gave us some suggestions. Their suggestions made us understand our strengths and weaknesses. We could reflect on our presentations and think how to improve ourselves. It also let us experience doing a presentation in front of others. We improved our performance on the stage. We learned extensively and widely, not just limited to the content of the textbook. (Student 7)

這是很棒的一個活動, 透過同學們報告, 讓我們為他評分, 評分的過程，也讓我們學習接受別人的意見，別人為我們評分時，會給一些建議，同學的建議可使我們了解自己的優缺點, 反省自己的報告並且去思考如何改進，也可以讓我們有上台報告的經驗，讓台風變得更好，也使我們學習更多、更廣，不再只有學習課本上的東西而已。

Three of the 69 students reported they disagreed. They did not learn English from PA (Item 2), but they did learn to give suggestions (Item 6). Since these groups of students only experienced this type of student assessing activity once, they might need practice of PA and SA before they would be able to identify the long-term improvement in their English abilities. Also, giving concrete suggestions is relatively more difficult than giving ratings and therefore needs more guidance.

Table 4 Informative function of reinforcement

	strongly agree	somewhat agree	neutral	somewhat disagree	strongly disagree	total
1. I paid more attention to my classmates' presentations when I evaluated them.	72.46% 50	24.64% 17	1.45% 1	1.45% 1	0.00% 0	100% 69
2. I learned English from evaluating my classmates' presentations.	55.07% 38	39.13% 27	1.45% 1	4.35% 3	0.00% 0	100% 69
3. I learned how to do a good oral presentation from rating my classmates.	78.26% 54	17.39% 12	1.45% 1	2.90% 1	0.00% 0	100% 69
4. My classmates' feedback was helpful to my presentation.	72.46% 50	26.09% 18	0.00% 0	1.45% 1	0.00% 0	100% 69
5. I could reflect my own presentation and think how to improve from evaluating myself.	73.91% 51	21.74% 15	2.90% 2	1.45% 1	0.00% 0	100% 69
6. I learned how to give clear concrete suggestions from	69.57% 48	24.64% 17	1.45% 1	4.35% 3	0.00% 0	100% 69

giving my classmates feedback.						
7. I learned how to encourage the presenter from giving my classmates feedback.	63.77% 44	30.43% 21	4.35% 3	1.45% 1	0.00% 0	100% 69

5.2.2 Motivational reinforcement

The majority of the students enjoyed being empowered to be assessors, and therefore they tried to fulfill the responsibilities of assessors and learn to be fair. In Item 8 and Item 9, the students reported they liked the assessing activity and they were able to assess their peers objectively (see Table 5). They knew they were playing the role of a teacher.

Table 5 Motivational function of reinforcement

	strongly agree	somewhat agree	neutral	somewhat disagree	strongly disagree	total
8. I liked this assessing activity.	56.52% 39	39.13% 27	2.90% 2	1.45% 1	0.00% 0	100% 69
9. I could assess my classmates objectively.	60.87% 42	34.78% 24	2.90% 2	1.45% 1	0.00% 0	100% 69
10. I could assess myself objectively.	53.62% 37	37.68% 26	1.45% 1	4.34% 3	2.90% 2	100% 69
11. My classmates could assess me objectively.	68.12% 47	21.74% 15	5.80% 4	4.35% 3	0.00% 0	100% 69

> When doing peer assessment, I felt like a judge because I could evaluate my classmates. (Student 27)
> 同儕評分時，我覺得我像個評審一樣，因為可以幫同學評分。

> I think peer assessment needs to be fair and just. We can’t favor a particular classmate because he is a friend. Peer assessment is also a process to test whether I can give ratings in the stance of a teacher, so I think this is a very good activity. (Student 40)
> 我覺得同儕評分一定要公平公正，不能因為他是自己的朋友而偏袒他，所以同儕評分是在考驗是否能以老師的立場去評分，所以我覺得這個活動很好。

As Author 3 mentioned above, she thought most students could assess their peers and themselves objectively whereas only a few of them could not. In Table 5, five students reported that they could not assess themselves objectively (Item 10), and three students reported that they disagreed with the statement that their peers assessed them objectively (Item 11). One student doubted the fairness of peer assessment and their group played safe by giving a restricted range of ratings for all of the presenters:

> I don’t oppose this activity, but honestly half of the class and a few more didn’t take giving ratings seriously. It was always the same students [in the group] doing ratings. Some of the students couldn’t get the standard, just like our group. We were terrible in assessing. We gave two thirds of our classmates 16 [out of 20 possible points]. Once the teacher said one presenter was good, they changed the rating to 18. Also, friends and enemies influenced ratings more or less (I am not sure whether my class has this problem or not). (Student 13)
> 我並不反對這項活動, 但老實說班上半數在多一點的人再打分數上有點隨便, 打分數時幾乎都是那幾個在打, 部分的人再打分數上找不到標準, 像我們那一組打分數有夠兩光，班上三分之二的人都１６分，有次老師說她不錯，他們就把１６分改為１８分，另外朋友和仇人多少影響分數（我還不知道班上有無這個習慣）。

5.2.3 Cognitive reinforcement

The majority of the students agreed whole-class discussion of evaluation criteria helped them understand how to prepare for their presentations (Item 12) and that they had opportunities to talk about these criteria in their groups (Items 13-16) (see Table 6). The within-group discussions provided them opportunities to cultivate rapport, improve presentation, and assess others accurately. As one student observed during the interview,

> I feel group discussion was a very good task because it could build rapport among group members. Most important of all, we could absorb each other’s opinions. That helped us do a better presentation. It could also help me to increase accuracy of my evaluation of others. So I think we should have more group discussions. It helped me and others improve our abilities. (Student 66)

我覺得各組討論是一件很好的事情，因為可以培養組員的感情，最重要的事，可以吸收別人的意見，讓報告更完整，還可以增加自己評判別人的精準，所以我覺得應該多做各組討論，讓自己也讓別人提升自己的程度。

Table 6 Cognitive function of reinforcement

	strongly agree	somewhat agree	neutral	somewhat disagree	strongly disagree	total
12. The whole class discussion of evaluation criteria helped me understand how to prepare my oral presentation.	57.35% 39	38.24% 26	2.94% 2	1.47% 1	0.00% 0	100% 69
13. Each member had chance to express their own opinions in group discussions.	76.81% 53	11.59% 8	1.45% 1	8.70% 6	1.45% 1	100% 69
14. My group members accepted each other's opinions in group discussions.	71.01% 49	18.84% 13	2.90% 2	7.25% 5	0.00% 0	100% 69
15. My opinions were accepted in group discussions.	60.87% 42	30.43% 21	2.90% 2	2.90% 2	2.90% 2	100% 69
16. I had accepted my group members' opinion in group discussion.	78.26% 54	20.29% 14	0.00% 0	1.45% 1	0.00% 0	100% 69

Through discussion, the students learned how to accept diverse opinions and to work together to decide on a rating as a group.

When we gave ratings through group discussion, we learned not to raise or lower the standard because of particular people. (Student 50)
透過組內討論幫同學評分，我們就可以學會如何不因對象而提高或降低評分標準。

> Sometimes everyone had different opinions. After discussion, we could give a rating that everyone was satisfied with. (Student 31)
> 有時候大家意見很不合，但經過討論後，就會討論出大家都滿意的分數。

However, some students did not learn how to participate in and conduct an effective group discussion. A few students reported not every member had a chance to express their opinions, and that some of them did not accept each other's opinions (Items 13-15) (See Table 6). As Student 38 said, "Some people didn't respect others' opinions. They didn't learn to how to work well with each other." [有人不尊重別人的意見，沒辦法學會合作。]

6 Discussion

The findings of a strong positive correlation between PA and teacher assessments and moderate positive correlation between SA and teacher assessments imply that PA has a positive impact on SA, similarly to what was suggested by Topping and Ehly (2001). In the combination of both PA and SA, challenges that appear in either PA or SA alone in the previous studies are overcome. Contrary to previous arguments that young learners are not able to evaluate themselves fairly due to subjectivity and age-related issues of under-development of cognition and wishful thinking (Butler & Lee, 2006; Ross, 2006), this group of learners has demonstrated that they were able to conduct PA and SA as their teacher did, at least to a moderate extent. The problems of over-marking and under-marking were minimized, as Dochy et al. (1999) argued, though subjective issues still appeared in a few SA cases and therefore should be emphasized in training.

As suggested in social cognitive theory, learning is regulated by interaction between external influence and self-directedness (Bandura, 1991). The integration of group PA and SA serves informative, motivational, and cognitive functions to reinforce students' learning to assess and assessing to learn (Bandura, 1991). For the informative function, the reflecting experience was amplified and had a positive impact on students in terms of being an assessor as well as a language learner. In this context which combined both PA and SA, the students observed their peers' performance in the perspective of an outsider whereas scrutinized their own performance in the viewpoint as an insider. The process to compare, contrast, and cross-check the perceptions of an outsider, an insider, and other outsiders crystalized the standard of each evaluation criteria for the students, who therefore benefited from the experience and developed the abilities to be assessors in both PA and SA. Meanwhile, attending to and reflecting on their peers' as well as their own presentations helped these students' future performances and English learning although it required more experience with student assessment to get the long-term effects of improving their English abilities (Butler & Lee, 2006). Also, the results suggest students need guidance to interpret feedback, so they can bridge the connection between feedback obtained and their work to improve their future performance (Sadler, 1998).

As to motivational function, playing the role of the teacher motivated the students to become a fair assessor. The concept of the authoritative role of teachers in Chinese culture empowered the students when they accepted ownership of classroom assessment, and this served as the best motivation to learn to assess fairly, just as a teacher did. Nevertheless, the traditional authoritative role of teachers played a double-edged sword. Other than inspiring the students to be competent assessors, the teacher's role affected the students' judgment of their peers' performance. One student indicated that his group had changed the score they had decided on in order to conform to the teacher' opinion. In other words, the teacher might still dominate the assessing process, and the teacher was likely to remain the only standard in the classroom. As the power of assessment was surrendered from the teacher to the students, and the classroom culture moved from teacher-center to student-center, the learners' tradition should be neither idealized nor neglected.

In terms of cognitive mediation, the students applied the evaluation criteria that they agreed with to evaluate and reflect on their classmates' performances and then to their own presentations. Students' familiarity with the criteria enhances the validity (Falchikov & Goldfinch, 2000). Furthermore, discussion within groups enabled the students to share opinions with each other and analyze their observations collaboratively. Peer-assisted learning has been found to foster social interaction and develop interpersonal skills (Topping & Ehly, 2001), but learning from collaboration should not be taken for granted. Students need help in carrying out exploratory talk to try out and re-organize ideas and therefore benefit from talking to learn (Wells & Wells, 1984).

It is also noteworthy that the incorporation of PA and SA helped the teacher to understand the students' learning and made the assessment more comprehensive than merely teacher assessment or either one of the student assessments. From PA, the perceptions of the majority of the students could be told from their grading, written comments, and oral feedback, all of which deepened the teacher's understanding of whether or to what extent the students knew the criteria of high-quality performance. SA revealed each student's own point of view regarding his or her performance and the effort put into the preparation of the performance. As the teacher pinpointed in her interview, not only the product but also the process should be valued, and she appreciated SA uncovered what she could not tell from the student's performance only.

The reciprocal nature of integrating PA and SA in the present study sheds light on the feasibility of implementing student assessment with young EFL learners. Being aware of students' traditional culture and avoiding romanticizing democratic practice of collaborative discussion empower every student, foster autonomy, and orient the learning and assessing process learner-centeredness.

References

Bandura, A. (1971). *Social learning theory*. New York: General Learning Press.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*, 248-287.

Birjandi, P., & Tamjid, N. H. (2012). The role of self-, peer and teacher assessment in promoting Iranian EFL learners' writing performance. *Assessment and Evaluation in Higher Education, 37*, 513-533.

Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning, 39*, 313-338. doi: 10.1111/j.1467-1770.1989.tb00595.x

Boud, D. (1995). *Enhancing learning through self-assessment*. London: Kogan Page.

Butler, R. (1990). The effects of mastery and competitive conditions on self-assessment at different ages. *Child Development, 61*, 201-210. doi: 10.1111/1467-8624.ep9102040554

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessments among Korean elementary school students studying English. *The Modern Language Journal, 90*, 506-518. doi: 10.1111/j.1540-4781.2006.00463.x

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*, 5-31. doi: 10.1177/0265532209346370

Chen, Y.-M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12*, 235-262. doi: 10.1177/1362168807086293

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, R. J., & Swerdlik, M. E. (2005). *Psychological testing and assessment: An introduction to tests and measurement*. Boston: McGraw Hill.

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments?

Active Learning in Higher Education, 13, 129-142. doi: 10.1177/1469787412441284

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*, 331-350. doi: 10.1080/03075079912331379935

Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing, 18*, 111-131. doi: 10.1016/j.asw.2012.12.002

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*, 287-322.

Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment: Reflections on use of tutor, peer and self-assessment. *Teaching in Higher Education, 6*, 229-246. doi: 10.1080/13562510120045212

Harris, M. (1997). Self-assessment of language learning in formal settings. *ELT Journal, 51*, 12-20. doi: 10.1093/elt/51.1.12

Isaac, S., & Michael, W. (1995). *Handbook in research and evaluation for education and the behavioral sciences* (3rd ed.). San Diego: Educational and Industrial Testing Services.

Hung, Y.-J., Chen, S.-C., & Samuelson, B. L. (under review). Peer assessment of oral English performance in a Taiwanese elementary school.

Kwan, K.-P., & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education, 21*, 205-214.

Landis, J. R., & Koch, G. D. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Langan, A. M., Shuker, D. M., Cullen, W. R., Penney, D., Preziosi, R. F., & Wheater, C. P. (2008). Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment and Evaluation in Higher Education, 33*, 179-190. doi: 10.1080/02602930701292498

Lanning, S. K., Brickhouse, T. H., Gunsolley, J. C., Ranson, S. L., & Willett, R. M. (2011). Communication skills instruction: An analysis of self, peer-group, student instructors and faculty assessment. *Patient Education and Counseling, 83*, 145-151. doi: 10.1016/j.pec.2010.06.024

Mok, J. (2010). A case study of students' perceptions of peer assessment in Hong Kong. *ELT Journal, 65*, 230-239. doi: 10.1093/elt/ccq062

Murakami, C., Valvona, C., & Broudy, D. (2012). Turning apathy into activeness in oral communication classes: Regular self- and peer-assessment in a TBLT programme. *System, 40*, 407-420. doi: 10.1016/j.system.2012.07.003

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education, 27*, 309-323.

Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Encyclopedia of Language and Education* (Vol. 7, pp. 175-187). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Panadero, E., Romero, M., & Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation, 39*, 195-203. doi: 10.1016/j.stueduc.2013.10.005

Pond, K., UI-Hag, R., & Wade, W. (1995). Peer review: A precursor to peer assessment. *Innovation in Education and Training International, 32*, 314-323.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research and Evaluation, 11*, 1-13.

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5-6 mathematics effects on problem-solving achievement. *Educational Assessment, 8*, 43-59.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy and Practice, 5*, 77-85.

Schunk, D. H. (2001). Social cognitive theory and self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Shen, C., Lin, F., Xu, Y., Guo, W., Guo, F., Chen, M., . . . , & Liu, S. (2001). *Enjoy 10*. Tainan, Taiwan: Bureau of Education of Tainan City Government.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249-276.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*, 20-27.

Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction, 20*, 339-343.

Topping, K. J., & Ehly, S. W. (2001). Peer assisted learning: A framework for consultation. *Journal of Educational and Psychological Consultation, 12*, 113-132.

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*, 270-279.

Wells, G., & Wells, J. (1984). Learning to talk and talking to learn. *Theory into Practice, 23*, 190-197.

Ye, X. (2001). Alternative assessment. In Y. Shi (Ed.), *English teaching and assessment in primary and middle schools* (pp. 42-73). Taipei, Taiwan: Ministry of Education.