# Self- and Peer-Assessment of Speaking

**Soo Hyoung Joo**[1]

## INTRODUCTION

The research on the second language (L2) speaking assessment has been predominantly concerned with formal proficiency tests. However, with a growing interest in learning-oriented assessment (LOA), more researchers are interested in learner involvement in speaking assessment (Blanche & Merino, 1989; Luoma, 2004; Ocarson, 1989). Self- and peer-assessment of speaking, where learners evaluate the performance of their own and their peers, can be an exemplar of assessment with its primary purpose on learning (Chen, 2006, 2008; Ibersson, 2012; Saito, 2008; Topping & Ehly, 1998). Depending on the purpose, self- and peer-assessment can take various forms: a questionnaire on general speaking ability, a learning log for metacognitive reflection, or a classroom activity where learners use the same rating criteria as their teachers (Bachman & Palmer, 1989; Chen, 2008, Rivers, 2001).

Although it is of great theoretical and pedagogical value to investigate the nature of learners' judgments of L2 speaking, the challenge lies in that learners are not only providers of feedback; they, in fact, are also receivers of feedback with a goal to learn (Adams, Nuevo, & Egi, 2011). Therefore, depending on the interest on one of the two conflicting but conflated roles of learners, rating and learning, two lines of inquiry on self-and peer-assessment emerged. When the focus falls on learners' role as raters, the examination on whether the learners' rating matches with the teachers' rating becomes the main interest. On the other hand, when L2 learning is concerned, the whole process of assessment is investigated with regards to the multiple factors that either promote or inhibit the perception and incorporation of feedback.

This paper aims to address both questions: (1) do learners have the ability to assess the oral performances of their own and their peers, and (2) what factors in self- and peer-assessment affect the enhancement of L2 speaking ability. In order to answer these questions, the remainder of the paper will review the theoretical framework of LOA, and use the guiding questions from the LOA framework to review the empirical studies on both planned and unplanned assessment. While these empirical studies focused on varying factors influencing feedback perception and incorporation, their results commonly indicated that effective feedback should put the learner and their perception into the center of consideration.

## LEARNING-ORIENTED ASSESSMENT

The research on self- and peer-assessment stemmed from a broader field of inquiry referred to as formative assessment, assessment *for* learning, or learning-oriented assessment (LOA). These notions commonly contend the use of assessment information to make beneficial

[1] Soo Hyoung Joo received her M.A. in Applied Linguistics from Teachers College, Columbia University and is currently teaching English as a Forieign Language in a South Korean public middle school. Her research interest includes second language assessment and technology integration in English classes.

changes in instruction and learning, in contrast to the traditional notion assessment *of* learning, which uses assessment information as a report with summative purposes (Boston, 2002; Cizek, 2010; Turner & Purpura, 2015). Since the term, formative assessment, was first coined in the realm of curriculum evaluation (Scriven, 1967), the notion of prioritizing learning in both instruction and assessment has been expanded to K-12 and higher education contexts as well as L2 classrooms (Carless, 2007).

The impact of learner-generated feedback on learning is one of the central themes in LOA. LOA takes multiple interrelated dimensions of L2 learning into account: proficiency, elicitation, instructional, sociocognitive, affective, contextual, and interactional dimension (Turner & Purpura, 2015). Framing the review of self- and peer-assessment with the guiding questions from each of these dimensions can provide a holistic understanding of the learner's internal and external factors that affect self- and peer-assessment. First, for the proficiency dimension, whether learners have the ability to rate speaking proficiency can be discussed, in addition to the impact of L2 proficiency level on their assessment behavior. Second, in consideration of assisting learners to elicit quality feedback, the guiding questions of the elicitation and instructional dimensions can also be asked: what types of questions should be asked in the rating rubric and how can training enhance the quality of feedback? Then, as an attempt to uncover the interrelated factors that contribute to or inhibit learning during the process of self-and peer-feedback, questions from the sociocognitive dimension can guide the investigation of how self- and peer-assessment contribute to learning: what reasoning skills do self- and peer-assessment require of the learner and what do they learn from the process? The affective and interactional dimension can guide the investigation of psychoaffective factors influencing learning: to what extent is the interaction involved the assessment task anxiety producing? Finally, the questions from the contextual dimensions can inform the sociocultural forces that influence the stakeholders' attitude on the self- and peer-assessment task. The following sections will attempt to address these questions pertaining to the multiple dimensions in LOA.

## RATING SPEAKING PROFICIENCY

Given that speaking assessment requires a profound understanding of assessment criteria, the learners' ability to accurately assess self or peers' oral performances has often been challenged. The fact that language is both the medium of performance and a target of assessment, as well as that construct variables are hard to tease apart, prove just a few of the challenges inherent to assessing speaking (Kim, 2006; Luoma, 2004; McNamara, 1996). Learners as raters face more difficulty than professional raters in terms of their 1) limited proficiency in the language 2) lack of anonymity 3) rating activities inseparable from classroom practices (Jafarpur, 1991; Saito, 2008). The debate on whether learners can or cannot assess speaking ability generated a number of studies quantitatively comparing self- or peer-ratings to external criteria, such as another test or teachers' judgment, as extensively reviewed by Blanche and Merino (1989), Ross (1998), and Saito (2008).

Upon reviewing fifteen empirical studies, Blanche and Merino (1989) stated that the accuracy of self-assessment was high, since the values of the correlation coefficients between teacher and self-ratings ranging from .5 and higher were not uncommon. Ross (1998), however, rejected this argument through his meta-analysis of 29 studies on speaking self-assessment. The studies were reported to have effect sizes (i.e., the degree of association between teacher and self-ratings) of a range of .09 to 0.79. He concluded that the wide range of the effect size across studies suggested that self-assessment of speaking is susceptible to extraneous variables. Also, the median of the effect sizes, .53, was lower than that of reading or listening, suggesting that learners might be relatively inept at estimating their own speaking ability. Saito (2008) also conducted a meta-analysis, but focused on peer-assessment studies instead. The effect sizes of the peer and teacher ratings in the four studies were calculated, and were examined for consistency using the chi-square test of heterogeneity. He also pooled the data of 264 participants' ratings from all studies deriving an effect size of .50, which collectively indicated that the strong correlation between L2 peer and instructor ratings were relatively consistent across studies.

While these reviews provide an overview of the relationship between the learners' rating and teachers' rating, variables pertaining to the proficiency dimension beyond the statistical measures should also be considered, because individual studies or classroom contexts greatly vary. Studies on how speaking proficiency can be defined (Bachman & Palmer, 1989), whether a specific construct variable, grammatical competence, or proficiency levels can induce different rating behaviors (Cheng & Warren, 2005; Fujii & Mackey, 2009; Lee & Chang, 2005) fostered an understanding of the proficiency dimension of self-and peer-assessment. As part of the larger study examining construct validity of communicative language ability (Bachman & Palmer, 1981, 1982), Bachman and Palmer (1989) used the *multitrait multimethod design* to investigate the construct of communicative language ability. Extending from Canale and Swain's (1980) framework of communicative competence, Bachman and Palmer (1989) defined communicative language in terms of three *traits*: grammatical competence (e.g., morphology and syntax), pragmatic competence (e.g., vocabulary, cohesion, and organization), and sociolinguistic competence (e.g., register, nativeness, and nonliteral language). These three linguistic traits were then correlated with the three different *methods*, or question types, of self-assessment: 1) ability to use trait (e.g., Can you organize a speech in English with several different ideas in it?) 2) difficulty in using trait (e.g., How hard is it for you to use names of well-known American people and places in your speech?) and 3) recognition of input (e.g., Can you tell how polite English-speaking people are by the kind of English they use?). As a result of the statistical analyses, high coefficient alpha across traits were observed. With all three traits having strong loadings to the general communicative language abilities, Bachman and Palmer (1989) concluded that the self-assessment could be a reliable and valid method for assessing communicative competence.

Furthering Bachman and Palmer's (1989) examination on communicative competence, Cheng and Warren (2005), Lee and Chang (2005), and Fujii and Mackey (2009) zeroed in on one of the components: grammatical competence. These studies commonly responded to the criticism that learners do not have sufficient linguistic competence to rate others' linguistic accuracy of oral performances, either in the context of planned or unplanned assessment. Cheng and Warren (2005) examined whether the reliability of the peer's ratings on grammatical competence, defined in their test as *accuracy*, differs from the ratings on other variables of speaking, namely *content* and *delivery*. Fifty-one first-year undergraduate students in Hong Kong, who studied English for academic purposes, participated in the peer-assessment of seminar and oral presentation through an extensive 14-week period. Seminar and oral presentation were each evaluated by three teachers and peers within the group of four to five students. The analytic rating scale included the criteria *accuracy* (e.g. accuracy and appropriate use of vocabulary, structure, and register), *content* (e.g. relevance and interest of the topic), and *delivery* (e.g. pacing and appropriate use of body language). In general, high reliability of the peer assessment in comparison to the teachers' assessment was observed. They reported that no noticeable scoring pattern difference was observed between the rating of *language use* in comparison to other variables, suggesting that peer assessment on accuracy in oral performances can serve as a reliable source of assessment.

Similar to Cheng and Warren (2005), Lee and Chang's (2005) study examined the planned self- and peer-assessment with seven upper-intermediate learners of Korean. Their initial research focus was to examine whether learner engagement in rating criteria design can enhance the validity of self- and peer-assessment. However, subsequent analysis called for attention to the variable, grammatical competence, interpreted as *language use* in the rating scale. Participants were asked to deliver a series of three oral presentations on Korean culture during the 5-week intervention. Peer-assessment was conducted simultaneously with the presentations, and self-assessment was conducted as homework by watching the recorded performances. During presentation 1, students were asked to comment on presentations in an open-ended format without any scoring form. These comments were coded under four criteria (i.e., content, delivery, organization, and language use), which revealed that twice as many entries for content and delivery were observed in comparison to organization and language. For presentations 2 and 3, the student-developed analytic rating scales were used and the descriptive statistics of the four criteria highlighted that the greatest discrepancy between self-and peer-rating and teacher's ratings was observed in *language use*; learners tended to overrate peer's performance on *language use*.

With the contrastive findings between Cheng and Warren (2005) and Lee and Chang's (2005) studies on planned assessment, Fujii and Mackey (2009) studied the same topic, feedback on accuracy, but by investigating spontaneous, unplanned feedback. Naturally emerging peer-feedback during a communicative task was analyzed. Eighteen adult Japanese learners of English engaged in two open-ended decision-making tasks in pairs: a survival ranking task and a homestay decision tasks. The instances where non-target-like utterances occurred were tallied and the types of feedback that followed these utterances were coded. A low percentage (7% and 13% respectively for each task) of peer feedback followed non-target-like utterances. Moreover, peer feedback mostly targeted errors in argument structures or lexical choices, but none focused on grammatical accuracy, or morphosyntactic errors, such as subject-verb agreement, tense, plurals or article use. However, once peer feedback was provided, learners highly utilized it; 62% of the peer feedback was followed by modified output.

The three studies investigating the same construct, grammatical competence, reached a contrastive conclusion. Cheng & Warren's (2005) conclusion that learners have the ability to assess linguistic accuracy was contradicted by in the low quantity (i.e., low frequency) and quality (i.e., underrating) of feedback on morphosyntactic accuracy in Lee and Chang (2005) and Fujii and Mackey's (2009) studies. One of the factors to account for this discrepancy might be the divergent proficiency level of the participants. While the participants of Cheng and Warren's (2005) study were advanced learners of English from a university where English is the official medium of instruction, the participants of Lee and Chang's (2005) study were intermediate learners of Korean. Lee and Chang (2005) explicitly attributed the lack of reliability to the learners' limited proficiency level. Fujii and Mackey (2009) also attributed the lack of feedback on morphosyntax to the limited proficiency of the learners who were high-intermediate learners of English. They contrasted the exemplar-based system (i.e., focusing on lexical items and formulaic chunks) with the rule-based system (i.e., focusing on grammatical rules) on language acquisition, and stated that learners have heavier reliance on the exemplar-based system (Ellis, 2000). In other words, less reliance on the rule-based system limited the learners from focusing on forms. Also, learners were more focused on the completion of the task, which did not necessitate them to distinguish subtle meaning differences.

Given that the proficiency level of learners, especially for those with lower proficiency, can affect learners' ability to provide feedback, many researchers have been interested in whether assistance and training in eliciting quality feedback can benefit learners. The subsequent section will review the body of literature on this topic.

## ASSISTANCE TO ELICIT QUALITY FEEDBACK

The study on the assistance for learners to elicit quality feedback has two main threads of interests: the rating scale and the training effect. Researchers investigated whether clear rating criteria can ensure the consistency in rating (Babaii, Taghaddomi, & Pashmforoosh, 2015; Ibberson, 2012), and whether the training of learners, who are essentially novice raters, can enhance the quality of feedback (Patri, 2002; Saito, 2008; Sato & Lyster, 2012). A very recent investigation on rating criteria undertaken by Babaii et al. (2015) examined the reliability of self-assessment before and after the assessment criteria were provided. Twenty-nine Iranian undergraduate students learning English audio-recorded three short monologues each on different personal topics. Students were then given a sheet with blank lines where they were asked to list the criteria for rating. Using these individually selected criteria, students self-assessed their own recordings. After forty days, they were provided with a scoring criteria devised by the teachers. Followed by a training session using sample responses, students were once again asked to self-assess their own recording, this time, using the given scoring criteria.

The self-assessment results before and after the provision of the scoring criteria was compared using the paired-samples t-test. As a result, significant difference was observed between the scores before and after the provision of the criteria. The categorization of the student-selected self-assessment criteria revealed that the students lacked the macro-level variables such as *organization* or *strategic competence*. After the provision of the criteria, the magnitude of the correlation increased from .73 to .90, since students were aware of previously disregarded criteria. Babaii et al. (2015) concluded that the provision of the scoring criteria deepened learners' understanding of the construct variables consisting speaking ability resulting in a significant increase in the agreement between the learners' and teachers' scores.

While Babaii et al. (2015) compared the rating behaviors with or without the rating criteria, Ibberson (2012) extended the investigation by comparing the learners' use of different types of rating scales: checklist and rubric. Ibberson (2012) attempted to explore the validity of those rating scales both devised with the Common European Framework of Reference for Language (CEFR) for spoken production. The checklist comprised of one statement per each CEFR level, where the participants, fourteen learners of English in a university in U.K., could simply check one of the two choices: *able to use* or *difficult to use*. The second was an analytic rubric based on the four construct variables (range, accuracy, fluency, and coherence) of spoken language as defined in the CEFR. Each CEFR level had a detailed description of the criteria for each variable. The participants were trained to use both types of rating scale to rate their own recording of a two to three minute monologue, and rating data was collected each week. By the fifth week, the agreement between the teachers' rating and self-rating reached 73.7% when using the detailed rubric, a considerably high level of agreement in comparison to 42.2%, the agreement rate when using the checklist. Ibberson (2012) suggested that the well-devised analytic rubric has a positive effect in generating self-assessment ratings as comparable to the teachers.

While Babaii et al. (2015) and Ibberson (2012) both focused on the impact of well-devised rating criteria, it should be noted that they commonly disregarded one potential confounding variable: training effect. Both studies provide training of how to use the rating criteria, which poses a challenge to distinguish whether results are due to the provision of the criteria, or the training that could have led learners to be better feedback providers. Thus, the studies that isolate training effect as an explanatory variable can be reviewed. This topic was investigated by Patri's (2002) quasi-experimental study and Saito's (2008) experimental research, which both examined planned assessment, as well as Sato and Lyster (2012)'s quasi-experimental study, which examined unplanned assessment.

Patri (2002) delved into how norming sessions can impact the reliability of the self- and peer-assessment. Fifty-six undergraduate students in Hong Kong, mostly from a remedial English class, were assigned to either a control or experimental group. While both groups conducted oral presentations and filled out the self- and peer-assessment rating sheet repeatedly throughout four weeks, only the experimental group had a norming session, which Patri (2002) refers to as the peer feedback session before conducting the assessment each week. During the norming session, learners shared their feedback on the presentations and compared their ratings with their peers. The rating data collected in the end of the five weeks were correlated with the teacher's rating. The experimental group had a correlation of .85 significant at the .05 level, which was higher than that of the control group, .49, which indicated that the repeated norming session induced a greater correlation between the teacher's rating and learners' rating.

Moving forward from Patri (2002)'s assertion that training helps learners have a better understanding of the rating criteria, Saito (2008) investigated the effect of training length. The study involved two sets of experiment with Japanese undergraduate students in a beginning level English speaking course. In the first experiment, both the treatment (*n*=37) and control group (*n*=37) received explicit instructions on the ideal quality of oral presentations, while, only the control group was involved in training and norming sessions for forty minutes. Since the correlation difference test suggested statistically insignificant difference between the two groups, the second experiment allowed for a longer training period for the experimental group, a total of 200 minutes throughout five weeks, to amplify the training effect. However, the second experiment also yielded insignificant qualitative differences between the two groups, a contrastive result to Patri's (2002) findings. Saito (2008) interpreted that explicit instruction on rating criteria given to both groups might have served as a sufficient condition for learners to obtain a similar "frame of reference" (p. 575) with that of the instructors, and further training might be unrealistic in classroom settings.

While Patri (2002) and Saito (2008) focused on the training effect in planned assessment instances where peer assessment was a focal task in and of itself, Sato and Lyster's (2012) quasi-experimental study investigated whether learners can be trained to provide corrective feedback when the focal task was a meaning-oriented peer-interaction activity. Four intact classes in a Japanese university were assigned different treatments. While the first class was trained to provide prompts (*n*=41), the second class was trained to provide recasts (*n*=46) while conducting the peer-interaction activities. The third class did not receive any training, but only engaged in the same peer-interaction activities, thus named the peer-interaction-only group (*n*=42). The fourth class was the control group. The training for the two classes had three stages that lasted for three weeks: modeling, practice, and use-in-context. First, two teachers modeled providing either prompt or recast to each other. Then, students practiced with a role-play scenario, where each interlocutor was given a different list of errors they had to incorporate in their stories and students practiced detecting the partner's errors and providing feedback. Finally, they used the type of feedback in an authentic context.

The data were collected through an individual picture-description task and a paired decision-making task. The picture-description task was used to measure both pruned (i.e., subtracting the fillers and pauses) and unpruned speech rate, a measurement of oral fluency. The decision-making task, only conducted during the post-test was used to examine the interactional pattern, such as whether feedback was given or modified output followed. The post-test data after the ten-week intervention suggested that the two trained classes displayed significant improvement in both overall accuracy and fluency. Whereas, the third class, peer-interaction-only group (*n*=42), outperformed the fourth class, the control group (*n*=38), only by fluency. Although recasts occurred more frequently than prompts, the performance difference between the prompt-trained and recast-trained classes was insignificant in both fluency and accuracy. Furthermore, both types of feedback were followed by a high percentage of modified output (85% of the prompts and 88% of the recasts), which demonstrated that the learners have developed the ability to autonomously monitor their speech. They also manifested a high frequency of self-modified output, where learners would self-correct their errors even when no feedback was provided.

Sato and Lyster (2012) concluded that learners "proceduralized their declarative knowledge" (p. 613). In other words, frequent opportunities to provide feedback and self-correct their errors turned the declarative knowledge (i.e., grammatical knowledge) to procedural knowledge (i.e., spontaneous production through automatized grammatical knowledge) during meaning-focused peer interaction. Being trained both as receivers and providers of peer feedback, learners developed the ability to notice errors in their classmates' speech, which was displayed in enhanced accuracy. Additionally, in contrast to the conception that feedback might hinder fluency rates, the speech rates of the feedback groups had insignificant difference from peer-interaction-only groups. Repetitive practice to retrieve grammar knowledge resulted in faster processing speed for accuracy while maintaining the fluency.

The studies on the effect of training show both consistent and inconsistent findings. Patri (2002) and Sato and Lyster (2012) commonly indicated that with well thought out assistance, students could make judgments on the speaking performances as comparable to those of the teachers. To the contrary, Saito (2008) pointed out that teachers should exert caution in implementing extensive training, since rating itself cannot be an ultimate goal in language learning. In accord with Saito's (2008) study, there are several tendencies of learners' rating, which remains consistent across studies with or without training (AlFallay, 2004; Cheng & Warren, 2005; Lee & Chang, 2005; Patri, 2002; Saito & Fujita, 2004): 1) Learners have a tendency to overrate their peers, often resulting in higher mean scores than the teachers' rating. 2) Learners tend to harshly assess themselves, resulting in an underrating. 3) Learners' ratings display a smaller standard deviation of the ratings in comparison to the teacher's ratings; they tend to refrain from providing grades in both low and high extreme.

These tendencies encapsulate the complexity of self- and peer-assessment, where the effect of cognitive, interactional, psychoaffective, and contextual factors might not be overridden by training or well-devised rating scales. Therefore, determining the quality of feedback solely with the match or mismatch of the teacher and learners' ratings might mislead to overemphasizing learners' role as raters and disregard the purpose of learning-oriented assessment- to consider learning as integral to assessment and instruction (Saito, 2008; Turner & Purpura, 2015).

With prioritization on learning rather than rating itself, investigation on whether self- and peer-assessment actually derive learning outcomes has been of interest in both planned and unplanned context. For instance, Ahangari, Rassekh-Alqol, and Hamed's (2013) experimental study with fifty-two graduate students in an Iranian university compared oral presentation scores between the control and experimental group. After six weeks, since it was only the experimental group that engaged in peer-assessment activity for the whole study, the independent t-test indicated that the two groups' scores were significantly different at the .05 level; the experimental group outperformed the control group. While Ahangari et al.(2013) hypothesized that peer-assessment could have promoted the understanding of their own strength and weaknesses in oral proficiency, the lack of empirical evidence, such as the learners' report or reflection on the assessment process, poses a challenge to approve or disapprove the claim. Similarly, the outcome of learning has been often examined in the realm of unplanned assessments as Fujii and Mackey (2009) and Sato and Lyster's (2012) reports on the percentage of modified output. However, whether learning took place or not cannot be solely determined by the investigation of modified output, since learners might be simply repeating the corrected input without the full understanding of the linguistic intent and meaning behind the feedback.

Therefore, the process of self- and peer-assessment, as well as the product, should be investigated in examining whether the feedback generated through self- or peer-assessment bridges the learning gap (Chen, 2008). The following section will review the body of research interested in the process of learning as well as the factors that influence learning.

## FACTORS INFLUENCING THE LEARNING PROCESS

From feedback to learning, multiple dimensions of sociocognitive, affective, interactional and contextual variables interact. With an increasing interest on the dynamic interplay between the learners' internal and external factors, interviews, observation notes, or reflection reports have been widely investigated to conceptualize the validity of self-and peer-assessment (Moss, 2003, as cited in Turner & Purpura, 2015).

With its primary interest in the sociocognitive dimension, Yoshida's (2008) comprehensive case study underscores the interaction of the multidimensional variables, including the psychoaffective and contextual factors. With a focal examination of a peer feedback setting, Yoshida (2008) investigated the factors influencing the understanding of feedback on morphosyntactic errors. She examined the interaction of three speakers of English learning Japanese during their pair work with the framework of expert-novice relationship. The data collected from the audio-recordings of the thirty hours of class sessions, observation notes, and a follow-up stimulated recall interview transcript were examined to see whether understanding of the linguistic content of the feedback was evident during interaction.

Yoshida (2008) incorporated Schmidt's (1995) concept of understanding in her study, and researched the factors influencing understanding in addition to the noticing of feedback. According to Schmidt, among the two levels of awareness associated with learning, noticing is a lower level and understanding accompanies a higher level of awareness. Therefore, Yoshida (2008) did not consider modified output as a sole criterion indicating learning, since even if modification is evident, understanding might not have taken place. The students' stimulated recall suggested that they were simply repeating the input provided by the peer without truly learning the nature of the errors.

Yoshida (2008) concluded that the proficiency level of learners themselves and their partners, and the affective aspects such as the pressure from interacting with higher proficiency learners influenced the understanding of feedback. In order to achieve understanding, collective scaffolding must occur, in which learners support each other and co-construct knowledge (Donato, 1994, as cited in Yoshida, 2008). The prerequisite for collective scaffolding is the establishment of intersubjectivity, the cognitive and social sharing of events and goals for task completion. However, according to the stimulated recall data, intersubjectivity was not established due to not only cognitive factors such as the knowledge about grammar forms but also the social and affective factors caused by the interaction between different proficiency levels. Yoshida suggested that in order to establish intersubjectivity, learners should be able to equally take the initiative in their interactions and be satisfied with their roles.

While Yoshida (2008) mainly considered identifying the interaction of factors influencing L2 learning, other studies situated their studies in either psychoaffective (Cheng and Warren, 1997; Chen, 2006; Lim, 2007) or contextual (Butler and Lee, 2010) dimensions for a deeper understanding. One of the earliest investigations on psychoaffective factors was conducted by Cheng and Warren (1997), who administered an attitude questionnaire to fifty-two intermediate learners of English. Learners were given the same questionnaire before and after the peer-assessment exercise in order to investigate whether the participation in the assessment changed the attitude of the learners. The questionnaire was comprised of four-items, to which the learners could either answer yes, no, or not sure. The analysis of the number of students who answered yes to each item suggested that students generally had a positive attitude on peer assessment both before and after the assessment (63.5%) but their anxiety in assessing their peers had been lowered through participation (from 48.1% to 21.2%). However, one notable pattern was observed in the response to the question, whether they felt that their peer-assessment was fair and responsible. Although the total percentage of those who responded yes slightly decreased (44.2% to 36.5%), a majority of the learners have changed their responses before and after the activity. Subsequent interviews showed that those who changed their perception to a negative direction often felt compelled to overrate their peers not to threaten the face of their peers, leading them to question the objectivity of the judgment.

Cheng and Warren (1997)'s questionnaire was later adapted by Chen (2006), who took a step further and used the self-reported data to infer potential learning benefits. 40 English major students in a Taiwanese university responded to the same questionnaire twice: during the first two-weeks of the intervention when training took place (pre-assessment) and after the self- and peer-assessment of the storytelling task (post-assessment). Consistent with Cheng and Warren's (1997) findings, the pre- and post- questionnaire analyzed through the paired t-test highlighted that the learners' perception generally moved towards a positive direction. Furthermore, the Student's t-test on the questionnaire statements asking for the opinion on learning benefits demonstrated a negatively skewed distribution. Also, a majority of students commented on the benefits in their written comments, which represented students' tendency to agree that self-and peer-assessment has learning benefits such as enhanced awareness of their strength and weaknesses, development of critical thinking ability, and the acquisition of oral skills.

Lim (2007)'s study that followed the same procedure as Cheng and Warren (1997) and Chen (2006) generally was in accordance with the previous studies in that learners reported to be more aware of their attributes of speaking and they felt more motivated by having an audience (i.e., their peer assessors). However, at the same time, some participants in Lim's (2007) study expressed their lack of confidence as a rater, especially for grammatical accuracy or pronunciation, even more when the peers they had to assess had a seemingly higher proficiency than themselves. This perception was contrastive to the actual high correlation with the teachers' ratings achieved in two-weeks with training; the Cronbach's alpha was .91 for self-assessment and .92 for peer-assessment.

The mixed disposition of learners' attitude on self- and peer-assessment can also be characterized by the learners' comments from the aforementioned studies. Patri (2002) illustrated the concerns from the learners, such as personal bias, friendship, and inconsistency in standards as the following:

> Different people said different things, don't know who was right.
> They cannot identify some of my mistakes and say I'm good.
> Some of them [peers] are too subjective. (p.125)

In contrast, Babaii et al. (2015) cited the learners' comments to highlight the perceived benefits for learning:

> I think it helps us to evaluate our own speaking ability and be able to fix our problems. I found out what errors do I make the most while speaking and what parts I should focus on to speak better. (p.12)

With an awareness of the influence of psychoaffective factors associated with self- and peer-feedback studies, Alfallay (2004) took a step further and examined whether these psychoaffective factors actually affected the quality of feedback. He attempted to investigate whether the traits of learners, such as motivation, self-esteem, and anxiety, can have an effect on the accuracy of rating. Using personality trait measurement instruments such as Attitude/Motivation Test Battery (Gardner, 1985) and Foreign Language Classroom Anxiety Scale (E. Horwitz, M. Horwitz, & Cope, 1986), he first analyzed the personality traits of the participants, seventy-eight learners of English in Saudi Arabia. Upon ranking the trait scores, the students with the highest and lowest 25% of scores were selected so as to contrast the rating behavior of the two extreme groups. Then, the participants were involved in the self-and peer-assessment of oral presentation, subsequently correlated with the teachers' assessment. The comparison of the correlation coefficient between the extreme two groups showed interesting patterns. While most groups tended to reliably assess themselves and their peers, integratively motivated students showed a higher correlation than the instrumentally motivated ones. Similarly, high achievers showed a higher correlation than the low achievers. Different from the conception, however, low self-esteem learners had a higher correlation than the high self-esteem group, which AlFalley (2004) attributes to their deliberate effort not to overrate. Although the criteria to measure the quality of feedback was limited to the correlation of teacher and student feedback, AlFalley (2004) was one of the first to conclude that self- and peer-assessment can be dynamically influenced by learner traits.

Although AlFalley (2004) viewed psychoaffective factor a trait internal to learners, the stakeholders’ traits cannot but be bound to the sociocultural factors. Hamp-Lyons’ (2007) notion of *exam culture* (i.e., focus on measurement or standard) in contrast to *learning culture* (i.e., focus on individuals’ learning progress) highlights how culture can frame the perception and interpretation of assessment. For instance, Cheng and Warren (1997) attributed the skepticism on the accuracy of feedback expressed by some of his learners to the cultural context in Hong Kong, a country of *exam culture*. Since learners were used to the authority of assessment associated with teachers’ feedback, they were reluctant to conceptualize learners’ feedback as useful to influence learning.

In light of the importance of the sociocultural context, Butler and Lee’s (2010) quasi-experimental study compared two public elementary schools from the same city in South Korea with contrastive socio-economic status (SES), to investigate whether the different social background may have an influence in attitude (i.e., anxiety, confidence, and motivation) towards the self-assessment activity. Among the four 5th grade English classes from each school, two classes were assigned as treatment groups, and the other two remained as control groups. During the intervention that lasted for a whole semester, students from the treatment group engaged in a unit self-assessment, conducted at the end of each unit, as well as summative self-assessment, conducted at the end of the semester. The influence of participation in self-assessment activities was evaluated with the pre- and post-test, and the oral component of the Cambridge Young Learners’ English Test (CYLE).

Linear regression analysis revealed that the participation on self-assessment activities had marginal but positive effects on the CYLE performance. However, the perception of learners did not differ across control and experimental group; a great discrepancy, in fact was observed between the two schools. The students from the higher SES school had significantly negative perception in comparison to the other school. The interview with the teachers suggested that the teachers’ perception was consistent with their students’ responses. The teacher from the high SES school reported that her students mostly receive English tutoring, thus the students’ ability to autonomously reflect on their own learning does not suit the expectations from the parents. In contrast, the teacher from the lower SES school had a positive attitude for incorporating the self-assessment activity in subsequent semester and thought of ways to improve the practice, such as sharing the purpose of the activity from the beginning of the semester and holding the students accountable for participation in the practice. Butler and Lee (2010) concluded that sociocultural factors as well as instructional pattern of the teachers could have effect on the stakeholders’ attitude and perception toward self-assessment, which may have potential impact on learning.

The investigation of the sociocognitive, pscyhoaffective, and contextual dimensions that could either promote or hinder learning exemplifies the multifaceted nature of self- and peer-assessment. With these factors in consideration, it becomes apparent that not all self- and peer-assessment can, in a true sense, be labeled as learning-oriented, since without the conditions met learning cannot occur. The following section will discuss suggestions for future studies to further investigate these conditions.

## DISCUSSION

In this paper, the multiple dimensions of LOA served as a guiding framework to inform the interrelated challenges associated with learners being a provider as well as the receiver of feedback. Overarching these challenges, all studies commonly agree that self- and peer-assessment is the most meaningful when learners and their learning is put at the center of consideration. With the understanding that some self- and peer-assessment activities can be more learning-oriented than others, I will discuss the four main themes that can be further investigated: controlling the quality of feedback, sharing learning evidence, exploration of different task characteristics, and greater integration of the assessment activity into the curriculum through technology.

Firstly, with the understanding that the provision of quality feedback is crucial to learning, self- and peer-feedback studies have employed different methods to control the quality of feedback. The correlation studies of teacher and student ratings mainly employed a rating scale as an instrument with the purpose of controlling extraneous variables. While a well-devised rating criteria can indeed guide learners to rate performance with consistency (Babaii et al., 2015, Bachman & Palmer, 1989; Ibberson, 2012), using rating scales in isolation can deprive students of opportunities to share open-ended comments, which could provide invaluable information conducive to learning. As Yoshida (2008) pointed out, collective scaffolding is a necessary condition to promote the understanding of feedback. This finding underscores the importance of providing learners with opportunity to collaboratively figure out the linguistic intent and mechanism through qualitative comments. The sharing of comments can take the form of an oral discussion as in Patri's (2002) norming session, or written feedback (Babaii et al., 2015; Chen, 2008).

Secondly, in a strict sense, assessment is considered learning-oriented only if the evidence of a system in the L2 change can be provided (Turner & Purpura, 2015). However, self- and peer-feedback studies did not attempt to empirically support the learning benefits assuming that self-and peer-assessment activities would raise learners' metacognitive awareness (Rivers, 2001), and given the positive comments by the learners (Chen, 2006; Lim 2007). Considering the contextual influence in exam cultures, where measurement drives perception (Cheng & Warren 1997; Butler & Lee, 2010), evidence of learning can influence the learners' attitude on self- and peer-assessment practices towards a positive direction.

Thirdly, given that different tasks require certain types of feedback (Luoma, 2004), further investigation on the relationship between task types and the students' assessment behavior can inform the design of the most suitable feedback for each task type. Studies focusing on planned assessment predominantly used oral presentation as the target of assessment, most likely due to the conception that formal speaking tasks benefit the most from metacognitive training (Ahangari, 2013; AlFallay, 2004; Lee & Chang, 2005; Patri, 2002; Saito, 2008). While assessment activities on formal speaking tasks indeed benefit the target language use domain for formal interactions, the frequent everyday conversations are predominantly informal and have characteristics non-transferrable from the formal speaking training. As unplanned assessment studies employed various tasks where the main purpose is to get the meaning across, such as storytelling or picture descriptions (Sato and Lyster, 2012; Yoshida, 2008), planned assessment studies should also investigate the use of these tasks, with the support of audio and video recording which are often already used in self-assessment studies. In addition, co-constructing task-specific criteria with learners could guide learners to have a deeper understanding of the task objectives and induce useful feedback on task specific performances (Lee & Chang, 2005; Luoma, 2004). However, dialogue tasks, in particular, should be implemented with caution, since learners might confound the interactional factors between the two interlocutors when providing feedback, a challenge that professional raters also face (Ducasse & Brown, 2009).

Finally, the stronger integration of self-and peer-assessment study into the curriculum can be attempted with the support of technology. In consideration of the two types of self-assessment, *performance-oriented self-assessment* (i.e., sampling a performance at one point in time) and *development-oriented self-assessment* (i.e., assessing changes over a period of time), which can be selectively chosen depending on the purposes (Bachman, 2000; Haughton & Dickinson, 1988; Oscarson, 1989, as cited in Saito, 2003), *development-oriented assessment* might be more suitable for learning-oriented purposes. As Babaii et al. (2015) and Lim (2001) collectively suggested, using self- and peer-assessment for summative purposes will be problematic given the learners' skepticism on the accuracy of feedback, regardless of the actual high correlation of ratings rendered by training. Instead, self- and peer-assessment activities would benefit learners the most when tightly integrated with the curriculum for learning-oriented purposes. Abraham, Stengel, and Welsh (2014) adapted peer-feedback as an integral component in a university Spanish course. With the assistance of technology, learners could easily connect to the web forum to provide feedback on not only language but also content for their final project. Similarly, Ko (2015) investigated the development and integration of a mobile application for peer-feedback on speaking in an EFL high school, which allowed more practice opportunities and promoted greater participation. Both studies underscored the potential of using technology to overcome the relative difficulties of providing feedback on oral communication, as well as to tightly integrate assessment into the curriculum (Luoma, 2004).

## CONCLUSION

The studies reviewed above suggest that when conditions are met, learners do have the ability to assess oral performances of their own and their peers. These conditions include: the clear provision of task-related criteria, sufficient training, considerations of the learner traits' and their perception, as well as the strong integration with the curriculum. Not only the product of assessment but also the process of being involved in self-and peer-assessment practices can enhance L2 speaking ability. While suggestions for further research have been provided, it should be noted that no definite model is available for self- and peer-assessment as a LOA, since the validity of LOA is highly context dependent. Thus, the consideration of multiple interactive dimensions can serve as a better guideline for self- and peer-assessment (Turner & Purpura, 2015).

## REFERENCES

Abraham, L. B., Stengel, P., & Welsh, S. (2014). *Affordances and constraints of technology-enhanced tools for learning-oriented assessment in second language learning*. Presentation at the Roundtable on Learning-Oriented Assessment in Language Classrooms and Large-Scale Contexts, Teachers College, Columbia University, New York.

Adams, R., Nuevo, A., & Egi, T. (2011). Explicit and implicit feedback, modified output, and SLA: Does explicit and implicit feedback promote learning and learner–learner interactions? *The Modern Language Journal, 95*(1), 42-63.

Ahangari, S., Rassekh-Alqol, B., & Hamed, L. A. A. (2013). The effect of peer assessment on oral presentation in an EFL context. *International Journal of Applied Linguistics and English Literature, 2*(3), 45-53.

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

Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2015). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing, 32*(3), 1-27.

Bachman, L. F., & Palmer, A. S. (1981). The construct validity of the FSI oral interview, *Language Learning, 31*(1), 67-86.

Bachman, L. F., & Palmer, A. S. (1982). The construct validity of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*(1), 14-29.

Bachman, L. F. (2000). Learner-directed assessment in ESL. In G. Ekbatani & H. Pierson (Eds.), *Learner-directed assessment in ESL* (pp. ix-xii). New Jersey, NJ: Lawrence Erlbaum Associates, Inc.

Blanche, P., & Merino, B. J. (1989). Self- assessment of foreign- language skills: Implications for teachers and researchers. *Language Learning, 39*(3), 313-338.

Boston, C. (2002). *The Concept of Formative Assessment*. (ERIC Digest No. 10). Retrieved from ERIC database. (ED470206)

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5–31.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics,1*(1), 1-47.

Carless, D. (2007). Learning- oriented assessment: conceptual bases and practical implications. *Innovations in Education and Teaching International, 44*(1), 57-66.

Chen Y. M. (2006). Peer- and self-assessment for English oral performance: A study of reliability and learning benefits. *English Teaching and Learning, 30*(4), 1–22.

Chen Y. M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12*(2), 235–262.

Cheng, W., & Warren, M. (1997). Having second thoughts: student perceptions before and after a peer assessment exercise. *Studies in Higher Education, 22*(2), 233-239.

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

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22*(1), 93-121.

Cizek, G. J. (2010). An introduction to formative assessment. In Andrade, H. L., & Cizek, G. J. (Eds.), *Handbook of Formative Assessment* (pp. 3-17). New York, NY: Routledge.

Donato, R. (1994). Collective scaffolding in second language learning. In J.P. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (p.33-56). Westport, CT: Ablex.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Testing, 26*, 423–443.

Ellis, R. (2000). Task-based research and language pedagogy. *Language Teaching Research 4*(3), 193-220.

Fujii, A., & Mackey, A. (2009). Interactional feedback in learner-learner interactions in a task-based EFL classroom. *International Review of Applied Linguistics in Language Teaching, 47*(3), 267-301.

Gardner, R., 1985. *Social psychology and second language learning*. London, U.K.: Arnold.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching. *International handbook of English language teaching* (pp. 487-504). Norwell, MA: Springer.

Haughton, G., & Dickinson, L. (1988). Collaborative assessment by masters' candidates in a tutor based system. *Language Testing, 5*, 233-246.

Horwitz, E., Horwitz, M., Cope, J., 1986. Foreign language classroom anxiety. *Modern Language Journal, 70*, 125–132.

Ibberson, H. (2012). Can Learners Self-assess Their Speaking Ability Accurately? *Multilingual Theory and Practice in Applied Linguistics*, 81-84.

Jafarpur, A. (1991). Can naive EFL learners estimate their own proficiency?. *Evaluation & Research in Education, 5*(3), 145-157.

Kim, H. J. (2006). Providing validity evidence for a speaking test using FACETS. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *6*(1), 1-37.

Ko, E. (2015). *Design, development, and evaluation of mobile application WikiTalki to promote English speaking skills in formal high school context* (Master's thesis, Ewha Womans University, Seoul, South Korea). Retrieved from http://www.riss.kr/link?id=T13818826

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

Lim, H. (2007). A Study of self-and peer-assessment of learners' oral proficiency. *CamLing Proceedings*, 169-176.

Luoma, S. (2004). *Assessing speaking*. New York, NY: Cambridge University Press.

McGroarty, M. E., & Zhu, W. (1997). Triangulation in classroom research: A study of peer revision. *Language Learning, 47*, 1–43.

Miller, L., & Ng, R. (1994). Peer assessment of oral language proficiency. *Perspectives: Working Papers in English and Communication, 6*(2), 41-56.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22*(4), 13-25.

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

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing, 19*(2), 109–131.

Rivers, W. P. (2001). Autonomy at All Costs: An Ethnography of Metacognitive Self-Assessment and Self- Management among Experienced Language Learners. *The Modern Language Journal, 85*(2), 279-290.

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

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research*, *8*(1), 31-54.

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553-581.

Saito, Y. (2003). The use of self-assessment in second language assessment. *TESOL Web Journal*, 3(1).

Sato, M., & Lyster, R. (2012). Peer interaction and corrective feedback for accuracy and fluency development. *Studies in Second Language Acquisition*, *34*(4), 591-626.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. *Attention and awareness in foreign language learning*, 1-63.

Turner, C. E. & Purpura, J. E. (2015). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Baneerjee (Eds.), *Handbook of Second Language Assessment*. Boston, MA: De Gruyter, Inc.

Topping, K., & Ehly, S. (Eds.). (1998). *Peer-assisted learning*. New York, NY: Routledge.

Yoshida, R. (2008). Learners' perception of corrective feedback in pair work. *Foreign Language Annals*, *41*(3), 525-541.