

The Interplay between Language Proficiency and Peer Assessment of Interactive Listening Tasks in EFL Classrooms

Angie H.C. Liu

Chung Yuan Christian University

angiehcliu@cycu.edu.tw

Abstract

In non-test situations, participants of interactive listening tasks often play the roles of co-constructors and evaluators simultaneously and yet most research studies on listening assessment ignore the aspects of learner involvement and peer assessment during the listening process. This study investigates the utility and reliability of peer assessment of interactive listening tasks in EFL classrooms and examines the impact of English language learners' (ELL) proficiency level on peer assessment outcomes. A total of 100 college ELLs in Taiwan were asked to evaluate the listening proficiency of their peers using a checklist of a five-point Likert scale immediately after completing three communicative tasks. The assessment criteria include global and local comprehension, strategy use, task success and discourse collaboration. The findings suggested that ELLs, without extensive training, achieved only moderate success in evaluating the listening performance of their peers when engaging in collaborative tasks. Nonetheless, ELL's language proficiency significantly affected the peer judgment such that ratings awarded by ELLs of high proficiency tended to be more reliable than those by ELLs of low proficiency. With regard to assessment criteria, higher agreement was detected on listening comprehension than listening strategies and discourse collaboration. Research and pedagogical implications were also discussed.

Key words: peer assessment, interactive listening, language proficiency, task-based assessment, test reliability

Introduction

The primary purpose of assessment is to collect information on the basis of examinees' test performance in order to make valid inferences regarding their performance in non-test situations. That is, the interpretations of test performance and uses of test scores should be validated for the intended purpose of the assessment (Bachman, 1990). While most listening tasks in real life involve interactive listening where the listener and the speaker co-construct the meaning of the exchanges and collaborate toward the goal of mutual understanding, most standardized tests of English language proficiency such as TOEFL, IELTS, TOEIC, GEPT and Cambridge ESOL examinations operationalize the assessment of listening proficiency through discrete, non-collaborative listening tasks. When taking these tests, English Language Learners (ELLs) interact solely with test materials and listening prompts without receiving any real-time feedback and mediation. Throughout the course of assessment, they also lack the opportunities to interact with other participants of the test.

To date, research studies on second language listening assessment have largely ignored the aspect of learner involvement during the listening process. Few studies examining the influence of interaction on language learning and evaluation incorporate the notions of cooperative learning and peer assessment in their research. However, even under the framework of peer assessment, research studies primarily investigate the validity and reliability of this measurement instrument on the basis of non-interactive listening tasks (Lazaraton, 1992, 1996; Ross, 1992; Ross & Berwick, 1992). Since interactive listening builds upon the co-construction of meaning among participants of test tasks, the assessment of it would logically involve the evaluative perception of the co-participants. During the process of reciprocal communication, both the speaker and the listener are required to engage in constant evaluation of each other's understanding of the transmitted information and make necessary modifications to the preliminary hypotheses on the basis of their judgments of the task success. However, such collaborative assessment of listening is rarely reported by second language testing research. Instead, most studies of collaborative assessment are found in the areas of science and math learning and development evaluation. Therefore, the purpose of this study is to investigate the utility and reliability of peer assessment of collaborative listening tasks in EFL classrooms. In view of the role played by the language proficiency on the effectiveness of peer assessment reported by other researchers (Falchikov, 1995; Kwan & Leung, 1996; Sullivan & Hall, 1997), this study also examines the impact of English language learners' (ELL) proficiency level on peer assessment outcomes.

Research Questions

Specifically, this study was designed to provide preliminary answers to the following questions:

1. To what extent can the peer assessment be used as a reliable measure of interactive listening proficiency of language learners in EFL classrooms?
2. What is the impact of EFL learners' language proficiency on the peer assessment outcome of interactive listening?
3. To what extent is the reliability of peer assessment affected by the rating criteria of interactive listening in EFL classrooms?

Review of Relevant Studies

In recent years, peer-assessment has received increasing attention owing to growing emphasis on learner-based language teaching and learner autonomy. However, studies on the reliability and validity of peer-assessments, to date, have suggested mixed results. A number of studies reported high correlations between results of teacher- and peer-assessment (Freeman, 1995; Hughes & Large, 1993; Miller and Ng, 1994; Rolfe, 1990) while others found contradictory results, contending that learners of low proficiency tended to

over-estimate their peers' language proficiency while learners of high proficiency tended to under-estimate (Jafarpur, 1991; Falchikov, 1995; Kwan & Leung, 1996; Orsmond et al., 1997; Sullivan & Hall, 1997).

Studies of peer-assessment, in general, suggest that the quality of peer-assessment is strongly linked to the understanding of performance standards, the descriptors used in the score scale, and the questionnaire questions used to elicit peer perceptions (Davidson & Henning, 1985; Heilenmann, 1990). Therefore, in order to increase the quality of peer-assessment, it is critical to provide detailed guidance and calibration training on the grading criteria to ensure that peer assessors clearly understand the goals of the evaluation tasks and reach uniformed interpretation of the performance standards and assessment criteria (Freeman, 1995; Smith & Smarkusky, 2005; Sullivan & Hall, 1997). Liu and Carless (2006) contended that requiring learners to actively engage in the process of clarifying the expected learning outcomes, monitoring and evaluating performance against the specified standards was crucial to the success of learner-centered assessment. Through explicit discussions about performance and standards and hands-on practice of peer evaluation, learners developed a more objective perspective in relation to expected performance level and thus were empowered to take an active role in the management of their own learning (Nicol & MacFarlane-Dick, 2006).

In investigating the influence of peer feedback on peer-assessment of oral skills, Patri (2002) examined the agreement among teacher-, self-, and peer-assessments of students in the presence of peer feedback and asserted that the peer feedback enhanced language learners' ability to judge their peers' oral presentation skills in such ways that were comparable to those of the teacher. Specifically, differences were found to exist between the peer-assessment outcome in the presence and absence of peer feedback. When peer feedback was present, the degree of agreement between peer- and teacher-assessments was relatively higher. He further suggested that language learners should be involved in the process of establishing rating criteria so that they can develop a deeper understanding of a successful performance.

Different kinds of assessment tasks elicit different types of interaction, which subsequently facilitates different types of learning (King, 2002). Therefore, it is essential to match the assessment tasks with the type of interaction that will facilitate the language learning that is intended to be promoted. Research on peer learning has shown that the interaction between and among the learners in a group greatly influences the intricacies of cognitive processing, which ultimately accounts for the learning that takes place (Cohen, 1994; O'Donnell & King, 1999; Webb & Palincsar, 1996). Specifically, group interaction comprising of information requests was found to promote comprehension skills while group interaction involving exchange of ideas, perspectives and opinions was found to facilitate high-level complex learning such as second language development. In order to promote high-level cognitive learning through group interaction, King (2002) proposed and

empirically validated a number of group-based interactive tasks such as “Pairs Squared”, “Peer-tutoring” and “Guided Reciprocal Peer Questioning”, which were structured and designed in specific ways to provide necessary support for the high-level cognitive processing to occur.

Studies published on the application of collaborative tasks in the assessment context are mostly in the area of science performance assessment (Bartlett, 1992; Lomask, Baron, Greigh, & Harrison, 1992; Neuberger, 1993; Shavelson & Baxter, 1992). When group-based collaborative assessment used in science learning evaluation, students are usually required to complete one part of the assessment on a group-discussion basis and other parts of the assessment on an individual basis. From the perspective of consequential validity, a major reason for designing group-based collaborative tasks is to align assessment practice more closely to the growing emphasis on small-group cooperative learning in classroom setting (Linn, 1993; Wise & Behuniak, 1993). The reported benefits of collaborative learning include the promotion of social skills, self-esteem, attitudes toward other learners, and student learning (Slavin, 1990).

Moreover, a number of studies have shown the great impact of group composition on group discussion quality and subsequently on student achievement regardless of assessment purposes. In investigating the effects of group ability composition on group processes and outcomes in science performance assessment, Webb et al. (1998) report that groups with above-average students produce more accurate and high-quality answers and explanations than groups without above-average students. As a result, below-average students who work with above-average students show higher achievement outcome. On the other hand, relatively poor performance was reported for below-average students who work without above-average students. In addition, high-ability students, when working in homogeneous groups, generally perform better than when they work in heterogeneous groups. In contrast, below-average students benefit more from staying in heterogeneous groups. The results of the study also confirm what have been reported by a number of previous studies of group composition and learning in classroom setting (Azmitia, 1988; Hooper & Hannafin, 1988; Hooper, Ward, Hannafin, & Clark, 1989) in that group ability composition has a major impact on performance and process of interactive, collaborative assessment.

The acquaintance between group members of collaborative tasks and its impact on the task performance was also studied by a number of second language testing researchers (Porter, 1991; O’Sullivan 2002). In the study conducted by Porter, he found no evidence to support the hypothesis of familiarity with one’s partner in an interactive task might positively affect performance. Contradictory to Porter’s finding, O’Sullivan reported evidence of an ‘acquaintanceship’ effect such that subjects achieved higher scores when collaborating with friends. The results of this study appeared to support the studies on second language acquisition, which suggested that learners modify their language when interacting with

speakers of various degree of familiarity (Plough and Gass, 1993; Tarone and Liu, 1995). However, analysis of the language indicated that there was no effect on linguistic complexity and that there was a sex-of-interlocutor by acquaintanceship interaction effect for accuracy.

Methodology

This section of the paper describes the following aspects of the study: characteristics of the participants, attributes of the interactive listening prompts, description of the peer-rating scale, study design, data collection, and data analysis methods.

Participants

A total of 100 college freshmen from two intact classes in Chung Yuan Christian University participated in this study, with 52 of them majoring in business administration and 48 of them in accounting. The average performance of the two classes on the Michigan English Proficiency Test is approximately comparable ($mean_{AC} = 55.3$ and $std_{AC} = 10.9$; $mean_{BA} = 56.4$ and $std_{BA} = 12.7$) and the level of English proficiency can be characterized as intermediate level.

Instruments and Materials

Interactive listening prompts. Three information-gap tasks were constructed as the collaborative listening prompts for eliciting genuine communication between the two parties of the pair discussion. The topics of the three tasks include 'which apartment to rent', 'mailing out Christmas presents in time', and 'what to order to dinner'. The information on the situational contexts and goals of the collaborative tasks was provided to both parties of the pair discussion. Each party received only partial yet complementary information regarding the collaborative tasks. In order to complete the tasks, both parties needed to exchange their information, discuss the scenario, and then reach a solution (see the appendix for examples of the information-gap tasks). Key words of the task information were supplemented with Chinese translation.

The peer-assessment rating scale. Participants' evaluation on the listening proficiency of their peers based on the interaction elicited by collaborative tasks was recorded on a five-point Likert scales (i.e., 1- strongly disagree, 2- slightly disagree, 3- neutral, 4- slightly agree, and 5- strongly agree). The rating scale consists of ten statements regarding various aspects of listening performance and the assessment criteria include global and local comprehension, strategy use, task success, and discourse collaboration. For example, 'my partner always understands the main message of the discussion', 'my partner always understands the details of the discussion', 'my partner always relies on my gestures to understand the discussion', and

‘my partner always needs to repeat what he/she said’. The overall reliability of the peer-assessment rating scale based on Cronbach’s alpha is .73, with the highest value for task 1 (.76), followed by task 3 (.74) and task 2 (.68).

Study Design and Data Collection Procedures

Participants were randomly divided into pairs and were asked to complete three information-gap tasks. The instructions for the collaborative tasks were provided in both Chinese and English to ensure the participants’ understanding of the procedures and objectives. Immediately after completing each collaborative task, participants were asked to evaluate the listening proficiency of their peers using the five-point Likert rating scale described above. Training on the use of the peer- assessment rating scale was provided prior to the assessment phase of the study.

Data Analysis

In order to evaluate the reliability of the peer-assessment as a scoring instrument for interactive listening proficiency of English language learners, the correlations between ratings of the three collaborative listening tasks were computed. The strength of the correlation across three interactive listening tasks was further analyzed based on the English proficiency level of the participants. Inter-task correlation coefficients were computed and compared across three levels of English proficiency (i.e., low, intermediate and high). Finally, inter-task correlations across the five rating criteria (i.e., global comprehension, local comprehension, strategy use, discourse collaboration and task success) were obtained in order to gauge their impact on the reliability of peer evaluating interactive listening ability of ELLs. Because the Likert scale used for peer assessment is an ordinal-level scale, Spearman’s rho correlation coefficients were chosen as the correlation statistics to report.

Results

Findings on the reliability of interactive listening tasks showed that the ratings awarded by college English language learners (ELLs) in Taiwan to their peers were moderately correlated ($r_{112}=.372$, $r_{123}=.417$, $r_{113}=.456$, $n=100$) and all correlations were statistically significant at the .01 level (see Table 1). That means, although the judgments on the college ELLs’ listening proficiency by their participating peers of the collaborative tasks were consistent across different tasks, the strength of that consistency was not very strong.

The influence of college ELLs’ English proficiency on the peer assessment outcome of interactive listening tasks was displayed in Table 2. It was shown that overall ratings awarded by college ELLs of higher proficiency were more reliable, with correlation coefficients ranging from .432 to .535, than those by learners of lower proficiency, with correlation

coefficients ranging from .297 to .484. However, when comparing the peer judgments of low- and intermediate-level of ELLs, it was found that ratings awarded by peer evaluators of low English proficiency, with correlation coefficients ranging from .297 to .343, were actually more consistent than those by peers of intermediate English proficiency, with correlation coefficients ranging from .297 to .343. Except for the inter-task correlations based on the peer evaluations of intermediate-level ELLs, all inter-task correlation coefficients were positive and statistically significant at the .05 or .01 levels. Nonetheless, task-specific variability was found to exist between peer evaluators of high and low English proficiency levels. For example, the correlations between tasks 1 and 3 and tasks 2 and 3 based on peer evaluators of low English proficiency ($r=.484$ and $.452$, respectively) were actually higher than that between tasks 1 and 2 based on peer evaluators of high English proficiency ($r=.432$).

Table 1
Overall Peer-assessment Consistency of Interactive Listening Tasks

	Task 1	Task 2	Task 3
Task 1 (correlation coefficient)	1.000	.372**	.456**
Sig. (2-tailed)	.	.000	.000
Task 2 (correlation coefficient)	.372**	1.000	.417**
Sig. (2-tailed)	.000	.	.000
Task 3 (correlation coefficient)	.456**	.417**	1.000
Sig. (2-tailed)	.000	.000	.

Note. ** indicates correlation is significant at the .01 level (2 tailed).

Table 2
Peer-assessment Consistency of Interactive Listening Tasks by English Proficiency Level

English Proficiency Level	Task1 * Task2	Task1 * Task3	Task2 * Task3
Low (correlation coefficient)	.362*	.484**	.452**
Sig. (2-tailed)	.049	.003	.005
N	33	33	33
Intermediate (correlation coefficient)	.297	.343	.309
Sig. (2-tailed)	.173	.097	.186
N	34	34	34
High (correlation coefficient)	.432**	.535**	.493**
Sig. (2-tailed)	.009	.000	.001
N	33	33	33

Note. ** indicates correlation is significant at the .01 level (2 tailed).

* indicates correlation is significant at the .05 level (2 tailed).

Results of the investigation on the extent to which rating criteria affect the reliability of peer assessment of interactive listening tasks in college EFL classrooms were displayed in Table 3. It was shown that except for the one between task1 and task 2 on the evaluation of task success, all inter-task correlation coefficients were positive and statistically significant at the .05 or .01 levels, regardless of the specific rating criteria employed in the peer assessment. In general, higher inter-task correlations were obtained on the peer judgments of listening comprehension attainment than those of listening strategy use (e.g. topic knowledge and key words), discourse collaboration (e.g., repair and turn-taking), and task success. Specifically, among the five rating criteria used in the assessment of various aspects of interactive listening performance, the overall strongest consistency across three collaborative tasks was found to occur in the assessment of local comprehension of the listening text ($r_{\text{mean}}=.685$, $n=100$), followed by the assessment of global comprehension of the listening text ($r_{\text{mean}}=.625$, $n=100$), the evaluation on the support received for discourse collaboration ($r_{\text{mean}}=.274$, $n=100$) and the use of cognitive listening strategies ($r_{\text{mean}}=.224$, $n=100$). The overall weakest consistency was found to occur in the judgments of the ultimate success of the collaborative tasks ($r_{\text{mean}}=.212$, $n=100$). At the same time, task-specific variability was identified with regard to the effect of rating criteria on the reliability of the peer evaluation of interactive listening

Table 3
Rating Criteria and Inter-task Correlations

Rating Criteria	Task1* Task2	Task1 * Task3	Task2 * Task3
Global Comprehension (correlation coefficient) Sig. (2-tailed)	.547** .000	.686** .000	.642** .000
Local Comprehension (correlation coefficient) Sig. (2-tailed)	.609** .000	.747** .000	.698** .000
Strategy Use (correlation coefficient) Sig. (2-tailed)	.243* .017	.231* .022	.198* .049
Discourse Collaboration (correlation coefficient) Sig. (2-tailed)	.232* .023	.279** .006	.312** .001
Task Success (correlation coefficient) Sig. (2-tailed)	.162 .109	.266** .009	.208* .036

Note. ** indicates correlation is significant at the .01 level (2 tailed).

* indicates correlation is significant at the .05 level (2 tailed).

ability of ELLs. For example, the correlations between tasks 1 and 3 and tasks 2 and 3 based on the peer evaluation of global comprehension of these listening tasks ($r=.686$ and $.642$, respectively) were actually higher than that between tasks 1 and 2 based on the peer assessment of local comprehension of these tasks ($r=.609$).

Discussion

The Dependability of Peer-assessment of Interactive Listening in EFL Classrooms

Since collaborative listening tasks require the active participation of both the listener and the speaker, it is reasonable to expect that the participating-members of the interactive events to be the most qualified evaluators of each other's listening proficiency as both parties are involved in the process of communication, during which oral texts from both sides are interpreted, monitored, evaluated and modified. However, the result of the investigation on the dependability of peer evaluation as an assessment instrument for interactive listening ability of college English learners in Taiwan indicated that college EFL learners in Taiwan, without extensive training, achieved only moderate degree of consistency in evaluating the listening performance of their peers when engaging in collaborative tasks.

There are a number of possible explanations for such a result. First, although peer-evaluators are privileged with the firsthand information about the interactive listening process, they lack the cognitive capacity to handle all the linguistic and nonlinguistic mechanism going on simultaneously during the real-time interactive listening, such as false starts, ungrammatical utterances, paraphrasing, repairs, turn-taking and non-verbal body language, facial expression and gestures. As participants of the two-way communication, they need to direct their full attention to the demands of the reciprocal collaboration and thus, are not able to put on the hats of the participants and the evaluators at the same time. Secondly, just like the assessment of all performance-based language tests, it takes calibration training and practice to produce consistent and dependable results, particularly when raters are learners of the language being evaluated. As suggested by other second language researchers, with extensive training and practice, it is likely that ELLs would develop better understanding of the mechanism involved in the interactive listening and the skills to direct their attention to the critical indicators of listening ability while ignoring nuisance performance irrelevant of the assessment of the listening construct. As a result, the reliability of peer evaluations on the listening ability of college ELLs in Taiwan through collaborative tasks could be enhanced. Thirdly, the moderate dependability of peer assessment outcomes could be related to the quality of the rating criteria used in this study. The overall reliability of the rating scale for this study is $.73$. It has been pointed out by researchers of previous studies that the quality of peer-assessment is strongly linked to the descriptors used in the score scale. If the reliability of the rating scale is improved, it is likely that the consistency of the peer evaluation of interactive listening tasks would increase as well.

The Interplay between Language Proficiency Level and Peer Assessment of Interactive Listening

This study found that the English proficiency level of the college students in Taiwan indeed significantly affected their judgments on the interactive listening ability of their peers such that ratings awarded by ELLs of high and low proficiency tended to be more reliable than those by ELLs of intermediate proficiency, despite of minor task-specific variability. Such findings provided additional insights to what was reported by earlier studies in that learners of low proficiency tended to over-estimate their peers' language proficiency while learners of high proficiency tended to under-estimate. What was added by this study is that the patterns of over-estimation and under-estimation of their peers' language proficiency were generally rather consistent for ELLs whose language proficiency was at the both ends of the ability scale. It could be argued that the quality (i.e., accuracy and dependability) of peer assessment is closely tied to the qualifications of peer evaluators. ELLs of higher proficiency levels are better equipped with the capability to consistently identify the failures and successes of their counterparts of the interactive process while ELLs of lower proficiency levels are consistently constrained by their own language deficiency to recognize the failures and successes of their counterparts. On the other hand, peer evaluators of intermediate proficiency sometimes over-estimate their peers' listening performance and other times under-estimate it. As a result, the reliability of their estimations fluctuated depending on the linguistic demands imposed by the communicative tasks and the subsequent performance of their counterparts of the interactive process.

The slightly more reliable peer assessment outcome produced by ELLs of higher proficiency than those of lower proficiency could be related to the intricate effect of group ability composition on the interactive listening process. Previous studies of group-based assessment concluded that group ability composition had a major impact on performance and process of interactive, collaborative assessment such that the performance and evaluation of students of higher ability were less affected by their group members than those of lower ability. Consequently, the judgments produced by peer evaluators of higher proficiency tended to be more consistent since they were also participants of the collaborative tasks. However, peer-evaluators of lower proficiency were more susceptible to the performance of their group members and thus, the judgment on their peers were relatively less stable. It is highly plausible that the degree of homogeneity between the language proficiency of the assessed ELLs and peer evaluators also contributed to the peer assessment quality of collaborative tasks, but this was not examined by this study.

The interplay between ELLs' language proficiency level and peer assessment of interactive listening was further complicated by the task-specific variability identified by this study. It appeared that in addition to the language proficiency of peer evaluators, the

reliability of peer assessment outcome was also affected by the characteristics of collaborative listening tasks. Until further analyses are conducted to investigate the ways task characteristics intertwine with the language proficiency of peer evaluators, the impact of ELLs' language proficiency on the reliability of peer assessment outcome cannot be fully understood.

The Impact of Scoring Criteria on the Reliability of Peer-assessment of Interactive Listening in EFL Classrooms

With regard to the assessment criteria used in evaluating the interactive listening ability of college ELLs in Taiwan, higher agreements were detected on the local and global comprehension of listening text than those on the use of listening strategies, deployment of discourse support, and the ultimate success of collaborative tasks. The differential correlations across various scoring criteria of interactive listening suggested that they posed different degrees of challenges for college ELLs in Taiwan when they were asked to conduct the evaluation task. The peer assessment on the global and local comprehension of collaborative listening text may be more reliable than that on the discourse collaboration and strategy use involved in interactive listening because the attainment of listening comprehension was meaning-based and required less inference-making while the evaluation on the deployment of discourse collaboration and use of cognitive listening strategies required the evaluators to recognize the occurrence of interactive behaviors and cognitive strategies such as turn-taking, repairs, clarification and paraphrasing. As for the ultimate success of collaborative listening tasks, the reason that the peer evaluation of this scoring criterion was least reliable could be due to the complexity of the communicative demands placed by the three information gap tasks, which require the ELL participants to first understand their portions of the written information provided, discuss with their partners regarding the objectives of the tasks and then verbally exchange and combine their portions of information to complete the communicative tasks. For all three information-gap tasks, there were multiple acceptable solutions / routes to complete the interactive tasks, which may very well add to the complexity and uncertainty on the part of the participating-evaluators to judge whether they had accomplished the goals of the tasks. Moreover, the higher reliabilities of the peer assessment of global and local listening comprehension attainment than those of discourse collaboration, strategy use and task success implied that while it may be appropriate to leave the assessment of some aspects of interactive listening tasks in the hands of peer evaluators, not all aspects of collaborative tasks are appropriate, especially when too much inference needs to be made.

Conclusions

In summary, the study results indicated that the peer assessment outcome of

collaborative listening tasks interacted with the language proficiency level of college English language learners in Taiwan in a complex way. The main findings include 1) college ELLs in Taiwan, without extensive training, achieved only moderate success in evaluating the listening performance of their peers when engaging in collaborative tasks, 2) peer judgments of the interactive listening ability of college ELLs in Taiwan were indeed affected by the language proficiency level of the peer evaluators, and 3) differential reliabilities were detected among various scoring criteria used in peer assessment of interactive listening tasks.

Testing is a means to collect performance information on test takers so that inferences can be made about their ability in non-test situations. Different kinds of assessment tasks elicit different types of test taker response, which subsequently facilitates different types of learning behaviors. Therefore, it is essential to match the assessment tasks with the type of response that will facilitate the language learning that is intended to be promoted. As listening comprehension in non-test situations is mostly an interactive process between the speaker and the listener, a valid and authentic listening assessment should also incorporate texts replicating natural spoken language and tasks resembling real-life situations with listening purposes explicitly stated. Moreover, through the act of peer-evaluation, ELLs are provided with the opportunities to engage in the tasks of analyzing, monitoring, and evaluating language process as well as language products and thus, benefit from the experience of linking assessment standards to language performance. However, despite of its appeal, the reliability of peer assessment of interactive listening tasks, as demonstrated by this study, is an intricate issue to tackle, as it is intertwined with the language proficiency of peer evaluators, scoring criteria and characteristics of the collaborative assessment prompts in a complex manner. To successfully implement interactive listening assessment in EFL classrooms, it is critical to untangle the complexity and obtain a better understanding of the parameters involved in the assessment process and outcome.

References

- Azmitia, M. (1988). Peer interaction and problem solving: when are two heads better than one? *Child Development*, 59, 87-96.
- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press: Oxford.
- Bartlett, L. D. (1992). Students successfully grapple with lessons of history in innovative group performance tasks. *Social Education*, 56, 101-102.
- Cohen, E. G. (1994). Restructuring the classroom: conditions for productive small groups. *Review of Educational Research*, 64, 1-36.
- Davidson, F. and Lynch, B. (2005). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. Yale University Press.

- Falchikov, N. (1995). Peer feedback marking: developing peer assessment. *Innovation in Education and Training International*, 32, 175-87.
- Freeman, M. (1995). Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education*, 20, 289-299.
- Heilenmann, K. L. (1990). Self assessment of second language ability: the role of response effects. *Language Testing*, 7, 174-201.
- Hooper, S., & Hannafin, M.J. (1988). Cooperative CBI: the effects of heterogeneous versus homogeneous grouping on the learning of progressively complex concepts. *Journal of Educational Computing Research*, 4, 413-424.
- Hooper, S., Ward, T.J. Hannafin, M.J., & Clark, H.T. (1989). The effects of aptitude composition on achievement during small group learning. *Journal of Computer-based Instruction*, 16, 102-109.
- Hughes, I.E. and Large, B.J. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education*, 18, 379-85.
- Jafapur, A. (1991). Can naïve EFL learners estimate their own proficiency? *Evaluation and Research in Education* 5, 145-57.
- King, A. (2002). Structuring peer interaction to promote high-level cognitive processing. *Theory into Practice*, 41(1), 33-39.
- Kwan, K. and Leung, R. (1996). Tutor versus peer group assessment of student performance in a stimulation training exercise. *Assessment and Evaluation in Higher Education*, 21, 239-49.
- Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing*, 13, 151-172.
- Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15, 1-16.
- Liu, N. and Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education*, 11(3), 279-290.
- Lomask, M., Baron, J., Greigh, J., and Harrison, C. (1992). *ConnMap: Connecticut's sue of concept mapping to assess the structure of students' knowledge of science*. A symposium presented at the annual meeting of the National Association of Research in Science Teaching, Cambridge, MA.
- Miller, L. and Ng, R. (1994). Peer assessment in oral language proficiency skills. *Perspectives: Working Papers in the Department of English*. City University of Hong Kong.
- Neuberger, W. (1993). *Making group assessments fair measures of students' abilities*. Paper presented at the National Center for Research on Evaluation, Standards, and Student Testing's Conference on "Assessment Questions: Equity Answers", UCLA, Los Angeles, CA.

- Nicol, D. & MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice, *Studies in Higher Education*, 31(2), 199-218.
- O'Donnell, A.M., and King, A. (1999). *Cognitive Perspectives on Peer Learning*. Mahwah, NJ: Erlbaum.
- Orsmond, P., Merry, S. and Reiling, K. (1997). A study in self-assessment: tutor and students' perceptions of performance criteria. *Assessment and Evaluation in Higher Education*, 22, 357-67.
- O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19, 277-295.
- Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, 19, 109-131.
- Plough, I. and Gass, S.M. (1993). Interlocutor and task familiarity: effects on interactional structure. In Crookes, G. and Gass, S.M. (eds.) *Tasks and Language Learning*. Clevedon: Multilingual Matters (pp. 35-56).
- Porter, D. (1991). Affective factors in language testing. In Alderson, J.C. and North, B. (eds.) *Language Testing in the 1990s*. Modern English Publications in association with the British Council. London: Macmillan (pp. 32-40).
- Rolfe, T. (1990). Self and peer-assessment in the ESL curriculum. In Brindley, G. (ed.) Vol. 6: *The Second Language Curriculum in Action*. Sydney: NCELTR, Macquarie University, 163-86.
- Ross, S. (1992). Accomodative questions in oral proficiency interviews. *Language Testing*, 9, 173-186.
- Ross, S. and Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14, 159-76.
- Shavelson, R. J., and Baxter, G. P. (1992). What we've learned about assessing hands-on science. *Educational Leadership*, 49, 20-25.
- Slavin, R. E. (1990). *Cooperative learning: Theory, research and practice*. Englewood Cliffs, NJ: Prentice-Hall.
- Smith, H. and Smarkusky, D. (2005). Competency matrices for peer assessment of individuals in tem projects. *Processing of SIGITE 05'* (pp. 155-162), Newark, New Jersey, USA.
- Sullivan, K. and Hall, C. (1997). Introducing students to self-assessment. *Assessment and Evaluation in Higher Education*, 22, 289-305.
- Tarone, E. and Liu, G.Q. (1995). Situational context, variation, and second language acquisition theory. In Cook, G. and Seidlhofer, B. (eds.) *Principle and Practice in Applied Linguistics*. Oxford: Oxford University Press (pp. 107-124).

*25th International Conference of English Teaching and Learning
2008 International Conference on English Instruction and Assessment*

- Webb, N.M., & Palincsar, A.S. (1996). Group processes in the classroom. In D.C. Berliner & R.C. Cafree (eds.) *Handbook of Educational Psychology* (pp. 841-873). New York: Simon & Shuster Mac-Millan.
- Webb, N.M., Nemer, K.M. & Chizhik, A.W. (1998). Equity issues in collaborative group assessment: group composition and performance. *American Educational Research Journal*, 35(4), 607-651.
- Wise, N. and Behuniak, P. (1993). *Collaboration in student assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Appendix

Task One: Renting an Apartment

Context and Goal of the Task: You would like to move out of the dormitory (宿舍) and into an apartment next semester. You have been checking out classified ads (分類廣告) for apartments for rent (出租公寓). Your best friend is helping you as well. Each of you has obtained different information. The goal of your task is to combine the information you've got and decide **which apartment would be the best choice for you and explain why.**

Part 1

Apartment A: on 7th floor; small room; no elevator; near night market (夜市); pets allowed; public bathroom; 10-minute walk to school; NT 3000

Apartment B: on 2nd floor; large room; with internet connection; 10-minute bus ride to school; public bathroom; no pets allowed; NT 3300

Apartment C: on 1st floor; with cable (有線) TV; no pets allowed; 20-minute MRT (捷運) ride to school; large room; public bathroom; NT 3000

Part 2

Apartment D: on 13th floor; with elevator; near swimming pool and sports center; 2-minute walk to school; pets allowed; small room; individual bath; NT 4000

Apartment E: on 3rd floor; near supermarket; individual bathroom; with air conditioning; 40-minute bus ride to school; pets not allowed; large room; NT 3000

Apartment F: on 1st floor; near a park; small room; 10-minute bicycle ride to school; public bathroom; kitchen available; pets allowed; NT 3200

Task Two: Christmas Presents

Context and Goal of the Task: Today is *December 11th*. The X'mas is coming up and you'd like to send gift packages to your friends overseas. Your partner and you have obtained postage and delivery information from three delivery companies. The goal of your task is to coordinate with each other to figure out the cheapest way to mail the four packages and ensure that they are received by X'mas. The sooner the packages get to your friends, the better. Please provide your answers to the questions below.

Questions

1. Which delivery company will you choose for each package?
2. How much does it cost for mailing all four packages?
3. When will all four packages be delivered?

Part 1: Package and Delivery Information

- Package 1 weights 4 pounds and needs to be sent to Japan.
- Package 2 weights 5 pounds and needs to be sent to Egypt.
- Postage and Delivery Chart

Company	Service Area	Fee Scale	Delivery Time
UPS	Asia, North America	(Asia) up to 3 pounds: NT 500; each additional pound: NT 300 (N.A.) up to 4 pounds: NT 800; 4~6 pounds: NT 1000	(Asia): 5 days (NA): 14 days
DHS	Africa, Europe	(Af) 300 NT per pound for up to five pounds (Eu) up to 2 pounds NT 500; each additional pound: NT 400	(Af): 8 days (Eu): 14 days
TPS	North America, Asia	(N.A.) up to 3 pounds: NT 650; each additional pound: NT 150 (Asia) less than 2 pounds: NT 300; 2~4 pounds: NT 550	(NA): 13 days (Asia): 4 days

Part 2: Package and Delivery Information

- Package 3 weights 6 pounds and needs to be sent to Canada.
- Package 4 weights 3 pounds and needs to be sent to Germany.
- Postage and Delivery Chart

*25th International Conference of English Teaching and Learning
2008 International Conference on English Instruction and Assessment*

Company	Service Area	Fee Scale	Delivery Time
UPS	Africa, Europe	(Af) up to 2 pounds: NT 400; each additional pound: NT 300 (Eu) up to 3 pounds: NT 1000; 4~6 pounds: NT 1800	(Af): 10 days (EU): 12 days
DHS	Asia, North America	(Asia) 300 NT per pound (N.A.) up to 2 pounds NT 500; 3~5 pounds: NT 900; each additional pound: NT 300	(Asia): 3 days (NA): 12 days
TPS	Europe, Africa	(Eu) up to 3 pounds: NT 800 (Af) up to 3 pounds: NT 400; 3~5 pounds: NT 1000	(Af): 14 days (Eu): 11 days

Task Three: What to Order for Dinner

Context and Goal of the Task: You and your partner are planning a small dinner party. Each of you received a list of *diet restrictions and preferences* (飲食限制及喜好) from your guests. You have obtained a menu list (菜單) from the restaurant of choice and need to decide what to order for the dinner party. You would like to order **three** entrees (主餐), **five** side dishes (附餐), and **three** desserts (甜點). The goal of your task is to coordinate with each other before finalizing the dinner menu to accommodate all guests. Please check mark () the final selections from the menu list.

Menu List

- I. *Entrée:* ___ Italian pasta in tomato sauce; ___ curry beef with rice; ___ sushi with cucumber; ___ hot pot with tofu; ___ Italian spaghetti with chicken; ___ fried rice with shrimps
- II. *Side Dish:* ___ mixed vegetables; ___ sweet corn; ___ mashed potato; ___ fish ball; ___ steamed egg; ___ baby carrots; ___ stinky tofu; ___ mushroom
- III. *Dessert:* ___ apple pie; ___ coffee cake; ___ banana pudding; ___ cheese cake; ___ fruit tarts

Diet Restrictions for Johnny, Megan, and Sophia (Part 1)

- Johnny is a vegetarian.
- Megan cannot have any caffeine.
- Sophia loves tofu.

Diet Restrictions for Peggy, Allen, and Kevin (Part 2)

- Peggy cannot have seafood.
- Allen loves Italian food.
- Kevin is allergic (過敏) to cheese.