Fall 2004

SLS 613

Instructor: J. D. Brown

Revising ELI listening/speaking diagnostic activities:

Developing rating scales for oral presentation and group discussion activities

Yusuke Fujisawa

ABSTRUCT

This project is an attempt to revise ELI listening / speaking diagnostic activities. After examining the current ELI listening/speaking diagnostic activities, I recognized that the current diagnostic activities were not giving feedback to students. Developing rating scales for oral presentation and group discussion activities was deemed useful for providing feedback to students. In consultation with the ELI instructor, analytical and task-specific rating scales were developed for oral presentation and group discussion activities. A pilot study was conducted to examine the applicability.

INTRODUCTION

Diagnostic tests are widely used in language programs. The purpose of such tests is to help students and their teachers focus their effort where they will be most effective (Brown, in press). Employing adequate diagnostic tests is thus an important part of language programs.

This project is an attempt to revise diagnostic activities currently employed in the English Language Institute (ELI) listening / speaking courses. In the first section, the purposes and the focus of the revision are explained. I examined the current ELI listening / speaking diagnostic activities which consist of oral presentation, lecture comprehension, and group discussion activities. After the examination, rating scales for oral presentation and group discussion activities were developed and found to be particularly useful as a part of revision. Thus, in the second section, the processes of the rating scales development are described. Finally, a pilot study was conducted in order to examine the applicability of the scales.

BACKGROUND

ELI Listening / Speaking Courses

The ELI is an English language program serving students who attend the University of

Hawaii (UH) and whose native language is other than English. Most are international students, but some are local immigrant students whose mother tongue is not English. Both graduate and undergraduate students enrolled in the ELI classes. The ELI provides instruction to help such students with academic English, effective study strategies, and integration into the UH-Manoa academic community, in order to facilitate their academic studies (ELI, n.d.a).

After being admitted to UH, all students who do not meet exemption criteria[1] must take the ELI placement test battery before registration. This test battery consists of two listening tests, two reading tests, and one writing test. Based on the scores, students are placed into intermediate or advanced courses, or are exempted in one or more of the three skill areas the test covers.

The current courses offered are as follows: an intermediate listening / speaking course (ELI 70), an advanced listening / speaking course (ELI 80), an intermediate reading course (ELI 72), an advanced reading course (ELI 82), an intermediate writing course (ELI 73), an advanced writing course for graduate students (ELI 83), and an advanced writing course for undergraduate students (ELI 100). The intermediate courses are followed by the advanced courses. That is to

---

[1] Exemption criteria are as follows: (1) the student is a native speaker of English, (2) the student has received a score of 600 or better on the paper-based TOEFL or a score of 250 or better on the computer-based TOEFL, (3) the student has received a score of 460 or better on the verbal section of the GRE, (4) the student has received a score of 460 or better on the verbal section of the SAT if taken before April 1995, or a score of 540 or better if taken in April 1995 or thereafter, (5) the student has an Associate of Arts degree from a community college within the University of Hawaii system, (6) the student has obtained the equivalent of 60 transferable semester credits with a GPA of 2.0 or better from a regionally accredited college or university in the United States, or the equivalent of 60 transferable semester credits with a GPA of 2.0 or better at a university whose academic standing is recognized by the University of Hawaii and where English is the primary language of instruction, and (7) the students has completed 6 years of full-time schooling with English as the medium of instruction at a middle school, high school, college, or university in Australia, Canada, (except Quebec), Ireland, New Zealand, the United Kingdom, or the United States. (ELI, n.d.b)

say, if one is placed in ELI 70, for example, one must take ELI 70 in the first semester and ELI 80 in the second semester.

The ELI listening / speaking courses (ELI 70 and 80), therefore, are designed to support students who are deemed to need supplementary English language course works for their academic success in terms of listening and speaking. Because the placement procedures are so carefully implemented, the misplacement of students rarely occur (K. Harsch, personal communication, November 19, 2004).

ELI Listening / Speaking Diagnostic Activities

The current ELI listening / speaking diagnostic activities were developed by Chang and Ito (2004) and are usually carried out in ELI 70 and ELI 80 during the first week of instruction. The purposes of the activities are (1) to help teachers assess their students' abilities in academic English in the area of aural comprehension and oral production, and (2) to support teachers' intuitions about students who might have been misplaced. In other words, the activities are designed for giving feedback to instructors and administrators.

Note that the activities are not mandatory. It is up to instructors whether or not to use the activities. This is part of the reason why they are called diagnostic activities rather than diagnostic tests (Y. Ito, personal communication, October 28, 2004).

The activities consist of two parts. Part 1 takes 30 to 35 minutes and part 2 takes approximately 40 minutes. Thus, usually instructors carry out part 1 and part 2 separately. Part 1 is an interview-and-presentation activity. There are four steps in this activity: (1) students form pairs and each student prepares a list of questions that he/she wants to ask a partner to get to know him/her, (2) one of the pair asks the questions of his / her partner and writes down the answers, (3) the other of the pair does the same thing, and (4) each student gives a short and

informal presentation about his / her partner to the whole class.

Part 2 consists of video-taped lecture comprehension activity and discussion activity. The procedures of part 2 are as follows: First, students are given a brief description of the content of the video-taped lecture. They are also given a question and answer sheet and are allowed to look over questions. Second, students watch the first fifteen minutes of video-taped lecture and answer the questions afterwards. Third, after all the students finished answering, they are divided into small groups and start discussing the question given based on the video-taped lecture. The lecture comprehension questions are composed of three sections. The first section tests summary writing and asks for the main idea, while the second and third section are testing short response and matching format and ask for more details (Chang & Ito, 2004).

In short, the activities are designed to assess the students' ability to perform three academic tasks: an oral presentation task, a lecture comprehension task, and a group discussion task. It is stated in the course objectives and goals of ELI 70 and 80 that these three tasks are the main focuses of ELI 70 and 80 (ELI, 2003). The results of needs analysis conducted in ELI 70 and 80 also support the legitimacy of emphasizing those three tasks in ELI 70 and ELI 80(Alexandrou & Revard, 1990; Park, Leong, & Hatcher, 1998; Kim, Kong, Lee, Silva & Urano,1999). As Brown and Hudson (2002) noted, direct relevancy of the content of diagnostic tests to the curriculum is an important characteristic of such tests. From that perspective, the current diagnostic activities are carefully designed.

However, the assessing procedures were not included or mentioned at all in the directions of the activities. No rating scales for scoring oral presentation and group discussion were provided. According to the informal interview with instructors, there is no common way to score student's performance on those activities. Thus, currently the activities were at most

used very intuitively for the instructors to get a grasp of students' oral and aural abilities. This is another reason why they are called activities rather than tests (Y. Ito, personal communication, October 28, 2004). Consequently, the possibility of giving systematic feedback has not been available.

Role of Feedback

The role of feedback and its importance have been widely discussed (e.g., Mendelsohn, 1991; Brown & Hudson, 1998, 2002).

Brown & Hudson (2002) explained three kinds of uses of feedback and its roles in language programs from the perspective of criterion-referenced tests. According to them, feedback may be useful for administrators, teachers, and students.

Feedback to administrators can help in many ways to shape and change any language curriculum. For instance, it helps "in making decisions about what the students really need to learn, which objectives are appropriate, how the test are working, which materials are effective, which teaching strategies are working well, and how well the curriculum is working"(p. 49).

For teachers, feedback can also be useful for understanding which objectives are not well-learned. In addition, feedback helps in identifying low achievers and high achievers in a class. Pointing out the low achievers on each objective may help the teacher provide extra help or materials to those students. Pointing out high achievers on an objective may also be useful because teachers can then give other materials on other objectives where those students need to work. Those high achievers may help other weaker students work on the objective in question as well.

For students, feedback as diagnostic information is particularly helpful. The information can help students in understanding which objectives are particularly weak and need

additional work, or which are particularly strong and need very little work. Such feedback may further help students set their personal goals in a program and become autonomous learners, which is one of the objectives of the ELI 70 and 80 (ELI, 2003; K. Harsch, personal communication, November 19, 2004).

In general, students are eager to know the results of any tests, and the feedback becomes an integral part of the learning process if the feedback is given in meaningful ways (Mendelsohn, 1991; Council of Europe, 2001; Brown & Hudson, 1998, 2002).

Council of Europe (2001) noted feedback was meaningful only if the recipient was ready to digest the information. Council of Europe (2001) summarized four such characteristics. First, recipients of feedback must be in a position to notice. In other words, recipients must be attentive, motivated, and familiar with the form in which the information is coming. Second, recipients must be ready to receive. Recipients cannot to be swamped with information and not have a way of recording, organizing and personalizing the information. Third, recipients must be ready to interpret. Recipients have to have sufficient pre-knowledge and awareness to understand the point of issue, and not to take counterproductive action. Fourth, Recipient must be in a position to integrate the information. In other words, recipients have to have the time, orientation and relevant resource to reflect on, integrate and so remember the new information.

In order to give such feedback, an adequate feedback instrument is necessary. However, Mendelsohn (1991) pointed out that "a weakness of most feedback in oral communication classes is the arbitrary, idiosyncratic, and unsystematic handling of errors" (p.199). He suggested that a systematic feedback instrument should be used, such as a rating checklist or a rating scale.

Thus, acknowledging the potential benefit of giving feedback to students, as well as to

instructors and administrators, and the importance of employing systematic feedback instrument, I came to conclusion that developing a systematic rating instrument for oral presentation activity and group discussion activity would be particularly useful in order to make the current ELI listening and speaking diagnostic activities more meaningful.

Two kinds of such systematic rating instrument are commonly used. One is a rating checklist. The checklist may be presented as a list of points like a questionnaire. Mendelsohn (1991), for example, proposed a rating check list for oral presentation and group interaction. According to Council of Europe (2001), the emphasis of the rating checklist is to show that "relevant ground has been covered, i.e., the emphasis is horizontal: how much of the content of the module has he/she successfully accomplished?" (p. 189). The advantage of using a rating checklist is that it allows raters to rate quickly.

The other kind of systematic rating instrument is a rating scale, on which the emphasis is on placing the person rated on a series of bands. Emphasis is vertical: "how far up the scale does he/she come?" (Council of Europe, 2001, p. 189).

Although it is attractive to use a rating checklist because of the practicality, I advocate a rating scale for this project because it enables us to give feedback as to the level of performance in relation to performance standards. One of the big concerns that students have in a classroom is whether the level of their performances is good enough or not. Only rating scales with performance descriptors can meet this need.

In addition, if used on multiple occasions, such rating scale may allow us to compare performances across different level of classes, that is, ELI 70 and 80. This enables us to understand not only the level difference between students in ELI 70 and 80, but also the progress that students make in the class. One of the instructors said, during the informal interview, that it

has been a concern of both students and instructors that no systematic indication of progress was provided in the class. Rating scales can meet this need as well. Thus, for this project, I decided to develop rating scales for oral presentation and group discussion activities.

RATING SCALES DEVELOPMENT PROCESSES

Concerns for Developing Rating Scales

Holistic vs. analytical scales. In developing a rating scale, the first concern is as to the type of rating scales. Broadly speaking, there are two kinds of rating scales: holistic scales and analytical scales. First, I will examine which type of rating scale is appropriate for the diagnostic activities under consideration.

Holistic scales require raters to judge their impression of a discourse according to its overall properties rather than providing separate scores for particular features of the language produced. In contrast, using analytical scales, a separate score is awarded for each of a number of features of a task. It has been reported that both holistic and analytical scales have advantages and disadvantages (e.g., Shohamy, Gordon, and Kraemer, 1992; Davies, Brown, Elder, Hill, Lumley, & MacNamara, 1999; Weigle, 2002; Luoma, 2004).

The biggest advantage that holistic scales have is the practicality. Holistic scales allow raters to make quick ratings because there is less to read and remember than in a complex grid with many criteria. However, it is often claimed that different raters may choose to focus on different aspects of the performance, leading potentially to poor reliability if only one rater is used. Also, holistic scales are not practical for diagnosing strength and weakness in individual student's performance.

In comparison with holistic scales, analytical scales have the following advantages. First, raters are required to focus on each of the nominated aspects of performance individually,

thus ensuring that they all are addressing the same features of the performance. Second, analytical scales allow for more exact and rich diagnostic information of specific strength and weakness in the student's performance, especially where skills may be developing at different rates. Third, analytical scales lead to greater reliability because each student is awarded a number of scores.

However, analytical scales may be problematic due to the possibility of a halo effect distorting the score, based on the number of judgments required. In other words, there is a possibility that raters may rate holistically and adjust analytical scores to match holistic impression. In addition, rating with analytical scales is time-consuming (Davies et al., 1999; Luoma, 2004).

Thus, holistic scales may be appropriate for large scale standardized tests due to its practicality. Yet, despite the disadvantages noted above, analytical scales seem to be more appropriate in this project because the main purpose of the activities is to give feedback to students as to their strength and weakness of the performance.

Task specific vs. universal. Another important point to consider is the degree of task-specificity of the rating scale. As Fulcher (1997) noted, at the one end is "a universal scale which is abstract and bear little relation to any context" (p. 79). At the other end, however, is "an attempt to construct a data base upon which description and explanation of performance can be modeled (and may be context bound in the sense of being relevant only to particular tasks or test takers)" (p. 80).

One can easily imagine that the former kind of scales are particularly suitable for a large scale oral proficiency test. Various people, in various contexts, take various tasks in such tests, so the rating scales must be as universal as possible. On the one hand, such rating scales enable

us to put a variety of students on the same scale, which is an important feature of norm-referenced test. The drawback of such rating scales, on the other hand, is that since it is designed to be universal, details are not measurable.

In this respect, the latter kind of task-specific rating scales are superior. Focusing on particular tasks and test takers, it is possible to measure the minute features of the performance. Thus, such rating scales are often recommended in a classroom evaluation (Mendelshon, 1991; Fulcher, 1997).

Considering that feedback to students would be more useful with detailed information, task-specific rating scales seem more appropriate for this project. Since the rating scales are to be used only in ELI 70 and 80 for assessing oral presentation and group discussion skills, the rating scales under consideration were specifically designed to describe the oral presentation and group discussion performances of students in ELI 70 and 80 in order to maximize the available information drawn from the results.

Adopting existing rating scales. Given that developing a rating scale from scratch is time-consuming and labor-intensive, it may be worth examining existing rating scales for oral skills. It is attractive to adopt well developed existing rating scales. Having said that, since holistic rating scales are not recommended for this project, holistic scales were not examined this time.

The result of the examination was rather disappointing, yet not surprising (See APPENDIX A for the summary grid). First, there were only a few analytical scales for speaking reported in the literature; second, most of these scales were for assessing oral proficiency rather than for specific performance on tasks. This is understandable, considering that the rating scales used in large scale tests tend to appear in the literature and holistic rating

scales are often used in such cases. As noted above, analytical rating scales require more time for rating than with holistic scales. In addition, because task-specific rating scales were not often applicable to other situations, it may be the case that such rating scales were not usually reported, except for research purposes such as Bonk and Ockey (2003) and Robinson (2003).

Even though the rating scales that were developed by Bonk and Ockey (2003) and Robinson (2003) were specifically designed to assess group discussion skills, which is the focus of this project, adopting these rating scales would still be problematic because the target population is quite different. Notice that these two rating scales were designed to assess Japanese university students who presumably have much lower English group discussion skills than the students in ELI 70 and 80. As the intention of the project is to develop a population-specific, as well as task-specific rating scales, it may not be appropriate to adopt these scales as are for this project.

Methodology for developing rating scales. Clauser (2000) and Luoma (2004) suggested some practical concerns when developing a new rating scale. Those include questions such as, "What aspects of the performance are to be scored?", "How many aspects are to be scored?", "How many levels are to be used?", and "What should the level descriptors say?" In order to answer those questions, various methodologies for developing rating scales have been proposed and researched (e.g., Fulcher, 1987, 1997; Chaloub-Deville, 1995; Shohamy et al., 1992; Upshur and Turner, 1995, and Norris, 2001). Council of Europe (2001) reviewed those methodologies and proposed three categories: intuitive methods, qualitative methods, and quantitative methods.

According to Council of Europe (2001), intuitive methods do not require any structured data collection but are based on principled interpretation of experience. The scale may be

developed by one person or in a group. Usually developers have considerable experience in teaching and assessing the students in the target population. They may consult existing scales, curriculum documents, teaching materials etc. The developer comes to share an understanding of the scale through repeated discussion, revision and scale application to learner performance until the scale is stabilized. Qualitative methods all involve small workshops with group of experts asking them to analyze data related to the scale. Those experts are asked to analyze either the level descriptors of the scale or sample performances at different levels. Quantitative methods utilize a considerable amount of statistical analyses.

Among those three types of methods, intuitive methods have been the most widely used methods (Luoma, 2004). Although qualitative or quantitative methods seem more rigorous, they are fairly time-consuming and require large amount of data. In this sense, strictly following qualitative and quantitative methods is not always feasible, especially in small scale studies as is the case of this project. Therefore for this project, I chose an intuitive method which I will explain in the next section.

Method for This Project

The procedures that I implemented had three steps: the first meeting, the second meeting, and the pilot study.

The first meeting. The attendees of the first meeting were the leading instructor of ELI listening / speaking courses and me. After providing a brief explanation of the project, levels to be included and aspects to be rated in the rating scale were discussed. The meeting lasted for 45 minutes.

As for the levels to be included, the more levels there are, the more specific the feedback will be, and the easier it will be to show progress. However, the lower the number of levels, the

more consistent the decisions will be. Luoma (2004) noted that "the compromise is somewhere in the middle, and test-specific scales often have four to six levels" (p. 80). We followed her suggestion and the levels were decided there were to four levels.

The aspects to be rated were then discussed separately for oral presentation activity and group discussion activity. To begin with, following the Council of Europe's (2001) suggestion, the number of aspects to be rated was set to be no more than five for both scales. The focus of the discussion was, "What aspects are emphasized in the ELI instruction?" As noted by Brown and Hudson (2002), it is important for criterion-referenced tests, including diagnostic activities under development, to reflect curriculum to the extent possible. In consulting the existing analytical rating scales, we came to the agreement that the following aspects were to be included in the rating scales: for oral presentation activity, fluency, intelligibility, vocabulary, organization/coherence, and delivery. For group discussion activity; fluency, intelligibility, vocabulary, communicative skills/strategies, and the discussion content. Some key features for each aspect were also identified through discussion.

After the first meeting, based on the discussion, I developed a draft rating scale by consulting existing rating scales. The level definitions were first developed. Then key features were defined for each aspect. Finally, descriptors were created for each level and for each aspect. The draft scales were then sent to the attendees of the second meeting along with the project overview.

The second meeting. The attendees of the second meeting were the leading instructor, two other ELI listening/speaking instructors and me. All of the instructors had the experience of teaching both in ELI 70 and 80.

Firstly, the project overview was examined. After the brief explanation of the project

and several questions and answers, the purposes and procedures of the project were deemed clarified. Secondly, the level definitions were discussed. The focus of the discussion was to examine whether these definitions are clear or not. Lastly, each descriptor was examined. Two questions were considered for each aspect: "Are the key features really key features?" and "Does each descriptor match to the typical performance of each level?" The meeting lasted for approximately 75 minutes.

Through the discussion, many comments were made. Among them, however, two issues were of particular concern. One was the issue of fluency. It was pointed out that the key features and descriptors of fluency were somewhat confusing and that the problem lay in the vague definition of fluency. The other concern was the issue of the oral presentation activity. We came to the agreement that the current oral presentation activity would not elicit enough performance for assessment with the rating scale because the presentation required in the activity was too short and informal.

After the second meeting, therefore, the definition of fluency was reexamined. In the literature, one thing all agree upon regarding fluency is that there is no one clear definition of fluency that is shared by all. Various authors defined fluency in their own way (e.g., Lennon, 1990; Schmidt, 1992; Fulcher, 1996; and Pawlikowska-Smith, 2002).

As summarized by Lennon (1990), the term fluency has been used in the broad sense and in the narrow sense. In the broad sense, the term fluency is used almost as a synonym for oral proficiency. In the narrow sense, however, fluency refers to "one, presumable component of oral proficiency" (p. 389). In this narrow sense, the emphasis seems to be on the temporal aspects of speech. Following the narrow sense, he suggested two key areas of performance that seem to be important for fluency: speech-pause relationships in performance and frequency of

occurrence of disfluency markers such as filled pauses and repetitions.

The oral proficiency rating scale adopted by The Canadian Language Benchmark 2000 is in line with Lennon's (1990) suggestion (Pawlikowska-Smith, 2002). In this rating scale the key features of fluency were deemed to be purely temporal aspect of speech such as speech rate, length of utterance unit, type of pauses, position of pauses, and frequency of disfluency markers, i.e., hesitation, repetitions, self-corrections, and false starts.

For this project, therefore, I decided to defined fluency from the temporal perspective only.

As for the issue of the oral presentation activity, I decided to adopt an alternative activity with minor a modification which is deemed to elicit enough performance samples. The original activity was developed by Ito (2000) as a classroom activity in the ELI 80. The modified procedures of the activity are as follows: (1) students form groups of three or four, (2) each of the group members chooses a different topic from an envelope, (3) for 10 minutes, students prepare for a three-minute presentation consisting of introduction, body, and conclusion, and (4) each member gives a three-minutes presentation in front of the group.

Pilot Study

Based on the discussion in the second meeting, the draft scales were modified. This modified draft rating scales were then piloted in a study. The purposes of the pilot study were (1) to examine whether the rating scales match with the level of students' performance, and (2) to examine if any confusion or difficulties occurred during the rating process. Accordingly, the research questions were as follows:

1. Are all the points in the rating scales used during the rating process?
2. Do there any confusion or difficulties occur during the rating process?

Participants. Participants in the pilot study were seven voluntary students from ELI 70 and 80. Three were from ELI 70 and four were from ELI 80.

Raters. Two graduate students from the Department of Second Language Studies were selected as raters. One was Japanese and the other was Korean. In order to avoid biasing the ratings due to previous knowledge about students, ELI instructors were not chosen as a rater.

Materials and procedures. The materials used for the pilot study were the oral presentation activity, that was adopted after the second meeting, and the group discussion activity (See APPENDIX B). The procedures for the oral presentation activity were already described above. The group discussion activity was carried out as follows: first, the content of the video-taped lecture was briefly introduced; second, students were asked to look through discussion questions for about one minute; third, the video-taped lecture was presented for 15 minutes; and fourth, students were asked to discuss the questions for 10 minutes. Students were allowed to take notes during the video-taped lecture. The topic of the video-taped lecture that was used for group discussion activity was the history of American Indian. This video-taped lecture has been used in Part 2 of the current diagnostic activities for the ELI 70.

The data collection was done for students in ELI 70 and 80 separately. Both procedures were, however, the same. Students were asked to do the oral presentation activity and group discussion activity consecutively. All the students' performances were video-recorded. The procedures lasted approximately one hour.

The raters were then asked to rate the video-recorded performances individually. Each rater rated all participants. The two ratings were then averaged. Raters were also asked to report any confusion or difficulties during the rating process.

Results and analysis. The descriptive statistics are provided in Table 1 below.

Cronbach's alpha of .95 for the oral presentation activity and of .87 for the group discussion activity was obtained, indicating that these activities were reasonably reliable.

Inter-rater reliability was then calculated for oral presentation activity and group discussion activity respectively. For the oral presentation activity, a strong inter-rater reliability was obtained ($r(5)=.943$, $p<.001$), indicating that rater 1 and rater 2 rated in a very similar way. For the group discussion activity, however, somewhat lower inter-reliability was found ($r(5)=.773$, $p<.05$), indicating that rater 1 and rater 2 behaved in a slightly different way.

Table 1
Descriptive statistics

Aspect	N	Minimum	Maximum	Mean	SD	Cronbach's alpha	inter-rater reliability
Pflu	7	1.5	4.0	2.79	0.86		
Pint	7	2.5	3.5	3.00	0.50		
Pvoc	7	1.5	4.0	2.71	0.91	0.95	0.94*
Porg	7	1.5	4.0	3.00	1.00		
Pdel	7	2.0	4.0	2.93	0.73		
Pave				2.89			
Dflu	7	2.0	4.0	2.93	0.67		
Dint	7	2.5	4.0	3.07	0.61		
Dvoc	7	2.0	4.0	2.79	0.70	0.87	0.77*
Dcom	7	2.0	4.0	3.21	0.76		
Ddis	7	2.0	4.0	3.07	0.67		
Dave				3.01			

Note: Pflu: presentation, fluency; Pint: presentation, intelligibility; Pvoc: presentation, vocabulary; Porg: presentation, organization/coherence; Pdel, presentation, delivery; Pave: presentation, average mean score; Dflu: discussion, fluency; Dint: discussion, intelligibility; Dvoc: discussion, vocabulary; Dcom: discussion, communication strategy; Ddis: discussion, discussion content; Dave: discussion, average mean score.
* significant at the .05 level.

The average mean score for the oral presentation activity was 2.89. Compared with the average mean score for the group discussion activity (3.01), raters gave somewhat lower rating for the performance on the oral presentation activity. All of the mean scores were between the ranges of 2.71 to 3.21 implying that no aspects were rated particularly leniently or harshly for

both activities.

The fact that the minimum scores for all aspects were all above 1.5 should be paid attention to because it indicates that the scale point 1 was not really used. This issue was further investigated.

Table 2 shows the frequency of the use of each scale point during the rating process. As can be seen from Table 2, most of the scale points were used for rating, indicating that the rating scales matched with the level of students' performance to some extent. Nevertheless, the scale point 1 was rarely used by both rater 1 and rater 2. As for the oral presentation activity, rater 1 did not use the scale point 1 for intelligibility. Rater 2, furthermore, did not use the scale point 1 at all. Rater 2 did not use the scale point 4 for intelligibility, either. From the group discussion activity, a clearer picture was obtained. Neither rater 1 nor rater 2 used the scale point 1 at all for all five aspects.

Table 2
The frequency of the use of each scale point

	Scale point	Pflu	Pint	Pvoc	Porg	Pdel	Dflu	Dint	Dvoc	Dcom	Ddis
Rater 1	4	2	3	2	3	2	1	1	1	2	1
	3	3	2	1	2	3	2	3	1	3	5
	2	1	2	3	1	1	4	3	5	2	1
	1	1	0	1	1	1	0	0	0	0	0
Rater 2	4	1	0	2	3	1	3	3	2	4	3
	3	3	6	2	1	5	3	4	4	2	2
	2	3	1	3	3	1	1	0	1	1	2
	1	0	0	0	0	0	0	0	0	0	0

As for the confusion or difficulties that raters reported, rater 1 pointed out that the amount of talk for the group discussion was not mentioned in the scale, but could be another key feature. Also, rater 1 noted that it was difficult to rate the quality of discussion content because it seemed strongly related to the range of vocabulary use. Sometimes it was the case that the idea seemed interesting and thoughtful, but the idea was not explained clearly due to a limited

vocabulary. Thus, rater 1 noted that the issue of whether or not the idea should be valued regardless of the clearness of explanation somewhat confused him. Rater 2 worried that her rating was inevitably affected a lot by the previous ratings. In other words, the ratings were done in a norm-referenced way, rather than in a criterion-referenced way which was deemed to be desirable in this case.

Discussion. The trend that the scale point 1 was scarcely used may be attributable to the fact that the pilot study was conducted at the end of the semester. The optimistic interpretation of the results, therefore, would be that the students' skills had improved so much through the semester that the students' performances do not correspond with the level which the scale point 1 described any more. If this interpretation is correct, the scale point 1 would be used if the scale was used at the beginning of the semester. However, this interpretation is somewhat optimistic.

The more realistic interpretation of the results would be that the level of descriptors for the scale point 1 was too low for the population. Since both rater 1 and rater 2 did not use the scale point 1 for all the five aspects of the performance of the group discussion activity, the level of descriptor for the scale point 1 was suspected to be too low for those aspects. Thus, those descriptors seemed to need modified.

Note that the descriptors for fluency, intelligibility and vocabulary were the same for both the oral presentation activity and the group discussion activity. Considering that at least rater 1 used the scale point 1 for those aspects of the performance of the oral presentation activity, the different descriptors for fluency, intelligibility, and vocabulary may be needed for the oral presentation activity and the group discussion activity.

Based on the implication drawn from the results of the pilot study, the draft rating scales

were further modified and finalized.

SUMMARY

This project was an attempt to add the function of providing feedback for students in addition to the current ELI listening/speaking diagnostic activities. The rating scales were, thus, developed for the oral presentation activity and group discussion activity through two meetings, with the help of ELI instructors. Some problems were identified and corrected through the pilot study. Thus, the finalized rating scales are expected to work reasonably well.

Last but not least, a pedagogical implication of this project should be mentioned. Although the rating scales were designed to be used for the ELI listening /speaking diagnostic activities, there are some other possible ways of using the scales. For instance, supposedly they could be used for other oral presentation and group discussion activities employed in the ELI classes as well. This could be valuable because, to my knowledge, no rating scale with descriptors for each level and aspect has been developed for the ELI listening / speaking classes. In addition, instructors can give the rating scale to students in advance of the activities to raise the students' awareness of the important aspects of the performance. I hope the rating scales will be used in beneficial ways for future instruction in the ELI.

REFERENCES

Alexandrou, A., & Revard, D. M. (1990) *A task-based need analysis for ELI listening courses.* Unpublished manuscript.

Bonk, W. J. & Ockey, G. J.(2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110.

Brown, J. D. & Hudson, T. (1998). The alternatives in language assessment. *TESLO quarterly*, *32*(4), 653-675.

Brown, J. D. & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University.

Brown, J. D. (in press).*Testing in language programs: A spreadsheet approach* (2nd ed., much revised). New York: McGraw-Hill.

Chaloub-Deville M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing 12*(1), 16-33.

Chang, S & Ito, Y. (2004). *Listening/speaking diagnostic activities.* University of Hawaii at Manoa, English Language Institute.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessment. *Applied Psychological Measurement*, *24*(4), 310-324.

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment.* Cambridge: Cambridge University.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & MacNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University.

English Language Institute (2003). *Goals and objectives: ELI listening & speaking.* University of Hawaii at Manoa, English Language Institute.

English Language Institute (n.d.a). *English Language Institute.* Retrieved November 25, 2004, from University of Hawaii at Manoa, English Language Institute Web site: http://www.hawaii.edu/eli/index.html

English Language Institute (n.d.b). *ELI clearance.* Retrieved November 25, 2004, from University of Hawaii at Manoa, English Language Institute Web site:http://www.hawaii.edu/eli/students/newstudents.html.

Folland, D. & Robertson, D. (1976). Toward Objectivity in group oral testing. *English Language Teaching journal, 30*, 157-167

Fulcher, G. (1987). Test of oral performance: the need for data based criteria. *English Language Teaching journal 41*(4), 287-291.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing,* 13(2), 208-38.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing*, *13*(1), 23-51.

Fulcher, G. (1997). The testing of speaking in a second language, In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Vol. 7, pp. 75-85). Dordrecht: Lluwer

Ito, Y. (2000). *Presentation practice in small groups.* University of Hawaii at Manoa, English Language Institute.

Kim, Y., Kong, D., Lee, Y., Silva, A., & Urano, K. (1999) *A task-based needs analysis for the ELI at the UH Manoa.* Unpublished manuscript.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387-417.

Luoma, S. (2004). *Assessing speaking.* Cambridge: Cambridge University.

Mendelsohn, D. J. (1991). Instruments for feedback in oral communication. *TESOL Journal*, Winter 1991-1992.

Norris, J. (2001). Identifying rating criteria for task-based EAP assessment. In T. Hudson & J. D. Brown, (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (Technical Report #21, pp. 163-204). Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.

Park, S. Leong, S. & Hatcher, J. (1998). *A needs analysis of oral proficiency at the ELI.* Unpublished manuscript.

Pawlikowska-Smith, G. (2002). *Canadian language benchmarks 2000: Theoretical framework.* Ottawa, Ontario: Canter for Canadian Language Benchmarks.

Robinson, P. (2003). *Aptitude-treatment interactions in the development of academic discussion ability by university-level Japanese L1 learners of English.* Paper presented at the Brownbag meeting at University of Hawaii at Manoa.

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies of Second Language Acquisition, 14*, 357-385.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of rater's background and training on the reliablity of direct writing tests. *The Modern Language Journal*, *76*, 27-33.

Upshur, J. A. & Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching journal*, *49*(1), 3-12.

APPENDIX A

Existing analytical rating scales for speaking

developer	intended use	target population	levels	aspects to be rated
Bonk & Ockey (2003)	group discussion	Japanese English-major university students	9	Pronunciation Fluency Grammar Vocabulary/content Communicative skills/strategies
Robinson (2003)	group discussion	university-level Japanese L1 learners of English	5	turn taking eye contact and gesture phrasal language discussion content
Council of Europe (2001)	proficiency	n/a	6	Range Accuracy Fluency Interaction Coherence
IELTS	proficiency	n/a	9	Fluency and coherence Lexical resource Grammatical range and accuracy Pronunciation
Canadian Language Benchmarks (2000)	proficiency	n/a	4 1- unable to achieve yet 2- needs help 3- satisfactory Benchmark achievement: pass 4- more than satisfactory achievement	Fluency Appropriateness Organization/ Coherence Vocabulary Grammar (accuracy) Intelligibility Relevance and adequacy of content Conversation management Negotiation of meaning
Ur (1996)	proficiency	n/a	5	Accuracy Fluency

APPENDIX B

The oral presentation activity and group discussion activity used in the pilot study

< Oral Presentation >

Instruction

In this activity, your task is:

- To present your opinions about the topic that you choose to others.

There are four steps to follow:

(1) 1 minute:	Get into groups of 3 or 4
(2) 1 minute:	Each of the group members chooses a different topic from an envelope.
(3) 10 minutes:	Prepare for a presentation (consisting of introduction, body, and conclusion).
(4) 3 minutes:	Each member gives a presentation in the group.

Your performance will be rated from the following perspectives:

- Fluency, Intelligibility, Vocabulary, Organization/coherence and, Delivery

< Group Discussion >

Instruction

In this activity, your task is:

- To listen to a videotaped lecture and discuss the content in a group.

There are four steps to follow:

(1) 1 minute:	Get into groups of 3 or 4
(2) 2 minutes:	Listen to the instructor's brief description of the content of the lecture and look through the discussion questions
(3) 15 minutes:	Listen to a videotaped lecture and take notes if you need. The lecture will be given only once.
(4) 10 minutes:	Discuss the questions in the group

Your performance will be rated from the following perspectives:

- Fluency , Intelligibility, Vocabulary, Communicative skills/strategies, Discussion content
-

Discussion questions

1. What is your image of American Indians? Where do you think this image comes from?
2. How would you describe the lecture's view of the American Indians in American history? Does her view differ from some of the stereotypical ideas you might have regarding the American Indians?
3. The early settlers decided that the Indians are human but it is O.K. to kill them. What do you think of their decision? If you were the early settlers, and the Indians were in the way of your owning land, what would you do?