Multi-faceted Rasch Analysis for the Assessment of
Group Oral Interaction Using CEFR Criteria

Junko NEGISHI
Open Education Centre, Waseda University

Abstract

Group oral test formats have recently been introduced to the range of oral performance tests because the assessment of L2 learners’ authentic conversational competence is considered important in the current era of globalization. However, these tests are not widely administered in Japan and only a few studies dealing with group oral interaction have been carried out to date. This study explored three facets - the severity of raters, the difficulty of rating categories, and participants’ speaking ability - when Japanese learners of English encountered oral interaction in a small group. The Common European Framework of Reference (CEFR; Council of Europe, 2001) was utilized for the rating in this study. The analysis was executed by means of multi-faceted Rasch measurement, which analyzed sources of variation and estimated “measures” from the raw scores given by raters. The raters were considered as self-consistent but showed significant differences among themselves as well as with other research. Utilizing multi-faceted analysis and building a consensus among raters by having them identify certain attributes in the speakers are recommended approaches to control such variances. The CEFR rating categories were proved to be valid; however, more detailed rating scales should be created for the situation in which all participants are Japanese, as the speaking ability of most of the participants was judged as low.

1. Introduction

1.1 Oral interaction in a group (group oral)

Paired and group oral test formats have recently been introduced to the range of oral performance tests because the assessment of L2 learners’ authentic conversational competence is considered important in the current era of globalization. The introduction of paired and group oral tests became possible because of “a product of the increased interpretability of test scores, potential validity of the scores when linked to real-world criteria, and positive washback effects” (Bonk and Ockey, 2003, p. 89). Oral performance tests of the paired or small group (group oral, hereafter) types are being administered, for example, in Cambridge First Certificate, Cambridge Certificate of Proficiency in English for paired and in the speaking test administered by the Council of Europe for both the paired and group oral. There
are some local tests that utilize the group oral in Korea given by the Educational Testing Service (ETS; to select scholarship recipients), in Hong Kong (Hong Kong Use of English) and in China (College English Test–Spoken English Test and Public English Test Systems). No group oral tests have been administered in Japan to date, excluding an English Forum for junior high schools organized municipally, whose purpose is to encourage and improve students’ communicative competence as advocated in the Japanese course of study. In comparison to interviews, the group oral is likely to produce natural and insightful conversation with peers, and it has been proved to be appropriate in certain test situations and in a battery of oral tests (Van Moere, 2006; Fulcher, 1996; Bonk and Ockey, 2003). Research dealing with the paired format has recently begun but only a few studies dealing with the group oral have been carried out to date.

It has been claimed that the group oral has various defects. First, Van Moere (2006) questions whether the test format allows interlocutors to demonstrate the full range of their linguistic knowledge. The second drawback is that the interaction might be affected by uncontrollable variables from other interlocutors. These variables may have a smaller effect on a single-speaker test format such as a monologue or picture description. The third limitation is a matter of the quality of the interlocutor. On certain occasions, when an interview is carried out, an interlocutor is a trained or professional interviewer who is supposed to elicit a required utterance from the test taker. However, an interlocutor in a group is also an examinee; therefore, the interlocutor cannot be trained and is less qualified.

Although the group oral has received little attention, it has various advantages. First, the group oral is resource economic (Bonk and Ockey, 2003). Second, raters can concentrate on their assessment. Controlling the conversation in an interview and assessing the interviewee are not easy tasks. This also means that rater training, since it only involves assessing group interactions, is less difficult than training an interviewer who is required to elicit a test-taker’s speech sample. Third, since the group oral promotes more communicative, speaking-focused teaching and learning goals, it may induce the washback effect (Hilsdon, 1995; Shohamy, Reves, and Bejarano, 1986). The co-constructed nature of interactions in conversation has been overlooked, as McNamara (1997) points out, and hence, the amount of support offered by the interlocutor should be considered. Finally, test-takers show positive reactions to the oral interactions in groups; they consider such discussions with other peers less intimidating than a face-to-face interview, in that they can control the conversation and are allowed to use more natural language (Shohamy et al., 1986; Fulcher, 1996). Bonk and Ockey conclude that “providing the students with opportunities to initiate and control conversation during the test might mean an enhancement of the validity of the score-based inferences” (p. 90).

1.2 Rater difference and multi-faceted Rasch analysis

Performance-based assessment can be subjective, and the scores awarded to test-takers may differ because they are based on human judgment. Since the actual assessment situation may produce interactions between the raters and candidates, task and candidates, and raters and task, there will be
great disparities in the scores. Rater difference is said to be the most variable because raters may be more severe or lenient towards a particular candidate, and they may interpret the rating scale differently or inconsistently. For these reasons, it is essential to conduct rater training in order to eliminate any conceivable disparity. Nonetheless, studies show that rater differences are still evident even after comprehensive training is conducted (e.g., Lumley, 2002; Lumley and McNamara, 1995; McNamara, 1996). With the differences in rater severity or leniency reaching as much as 40%, McNamara suggests that “a different approach is... to accept variability in stable rater characteristics as a fact of life, which must be compensated for in some way, either through multiple marking and averaging of scores, or using the more sophisticated techniques of multi-faceted analysis” (p. 469). The multi-faceted Rasch measurement is the method used to analyze sources of variation and estimates “measures” from the raw scores given by the raters.

1.3 The rating scale: CEFR

The Council of Europe created the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR) in 2001. This method of assessment has been actively used within the multi-linguistic and multi-cultural sphere of the European Union (EU). As part of the speaking test, the CEFR includes rating scales for group oral and pair interactions. It was beneficial to use the CEFR criteria because the Council of Europe disclosed all information such as its rationale and framework as well as the rating scales and a training DVD for raters. More importantly, candidates of the CEFR range from Basic Users to Proficient Users, which the former includes the participants of the current study, whereas other examinations introduced above (see 1.1) are administered for speakers above upper-intermediate to proficient level. However, the CEFR was mainly developed for assessing examinees within the area of the EU. This study, then, investigated whether or not the criteria were feasible for assessment in a situation where both the examinees and raters were Japanese.

The CEFR criteria are based on two types of rating scale: the CEFR Global Oral Assessment scale, which is considered to be a holistic rating scale, and the CEFR Oral Assessment Criteria Grid, which is an analytical scale consisting of five sub-categories: range, accuracy, fluency, interaction, and coherence. The criteria include a list format, that is, “can-do” descriptors consisting of six fundamental levels: A1 and A2 for Basic Users, B1, B2 for Independent Users, and C1, C2 for Proficient Users.

2. Research Questions

The aim of the study is to investigate the responses to the following three questions when Japanese learners of English encounter a group oral:
1) How reliable or internally consistent are the Japanese raters in terms of overall severity/leniency?
2) What and how consistent are the relative difficulties of the CEFR rating criteria/scales?
3) What are the participants’ speaking abilities according to the assessment of the raters?
3. Method

3.1 Participants

The participants in the study were 135 students. They were divided into a total of 45 groups, each containing three students. The groups comprised fifteen junior high school student groups, fifteen senior high school student groups, and fifteen university student groups. From the questionnaire distributed at the time of the group oral, we knew that no students had received education abroad with English as the medium of instruction.

3.2 Data collection

The data on the group oral were collected from each educational institution through the following process: (1) A questionnaire was distributed with questions on the participants’ backgrounds in terms of English language study; (2) The students were randomly allocated into groups of three; (3) Each group drew a card on which one of the six interaction topics - School, Family, Friends, Hobbies, English, and Culture (the last being only for university students) - was written down, and they were asked to speak on the topic; (4) Five minutes were allotted to each member of the group to plan his/her speech without speaking to the other members of the group; (5) Each member of the group introduced themselves for about half a minute as a warm-up activity; (6) Finally, the three students interacted orally as a group for five minutes on the selected topic. The interaction was videotaped to make DVDs for rating purposes.

3.3 Rating procedure

The ten raters who participated for the study were all Japanese teachers of English who held a minimum of a master’s degree in the field of English education or applied linguistics, and had been teaching English at either high schools or universities. While there are many reports of native speakers of English assessing Japanese learners of English, in this study, Japanese teachers rated the Japanese students since this is what most commonly occurs in reality.

The raters assessed all the participants’ group oral by watching the DVD on individual PCs. They rated the interaction using the seven levels of the CEFR - Below A1, A1, A2, B1, B2, C1, C2 - on both the Global Oral Assessment Scale and Oral Assessment Criteria Grid, in all the five sub-categories.

3.4 Multi-faceted Rasch analysis

For the analysis, this study utilized FACETS Version 3.6.4.0, a computer program for multi-faceted Rasch measurement (Linacre, 2008), which allowed us to transform raw scores obtained from the raters’ ratings into interval scales named “measures.” This study analyzed three facets - the severity of raters, the difficulty of rating categories, and the participants’ speaking ability.
4. Results

4.1 Rater measurement report

Table 1 Rater Measurement Report for Ten Raters

<table>
<thead>
<tr>
<th></th>
<th>Raw scores</th>
<th>Adjusted scores</th>
<th>Severity (logits)</th>
<th>Standard error</th>
<th>Infit mean square</th>
<th>Estimated discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater A</td>
<td>1.5</td>
<td>1.51</td>
<td>0.57</td>
<td>0.07</td>
<td>0.62</td>
<td>1.36</td>
</tr>
<tr>
<td>Rater B</td>
<td>1.7</td>
<td>1.69</td>
<td>-0.13</td>
<td>0.07</td>
<td>1.55</td>
<td>0.44</td>
</tr>
<tr>
<td>Rater C</td>
<td>2.2</td>
<td>2.05</td>
<td>-1.97</td>
<td>0.07</td>
<td>1.39</td>
<td>0.77</td>
</tr>
<tr>
<td>Rater D</td>
<td>1.9</td>
<td>1.87</td>
<td>-0.94</td>
<td>0.07</td>
<td>0.59</td>
<td>1.40</td>
</tr>
<tr>
<td>Rater E</td>
<td>1.5</td>
<td>1.56</td>
<td>0.37</td>
<td>0.07</td>
<td>1.03</td>
<td>0.93</td>
</tr>
<tr>
<td>Rater F</td>
<td>1.7</td>
<td>1.70</td>
<td>-0.17</td>
<td>0.07</td>
<td>0.91</td>
<td>1.10</td>
</tr>
<tr>
<td>Rater G</td>
<td>1.2</td>
<td>1.19</td>
<td>1.77</td>
<td>0.07</td>
<td>0.76</td>
<td>1.25</td>
</tr>
<tr>
<td>Rater H</td>
<td>1.4</td>
<td>1.37</td>
<td>1.06</td>
<td>0.07</td>
<td>0.93</td>
<td>1.16</td>
</tr>
<tr>
<td>Rater I</td>
<td>1.8</td>
<td>1.77</td>
<td>-0.47</td>
<td>0.07</td>
<td>1.51</td>
<td>0.77</td>
</tr>
<tr>
<td>Rater J</td>
<td>1.7</td>
<td>1.69</td>
<td>-0.99</td>
<td>0.07</td>
<td>0.71</td>
<td>1.31</td>
</tr>
<tr>
<td>Mean</td>
<td>1.66</td>
<td>1.64</td>
<td>0.00</td>
<td>0.07</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>S.D.</td>
<td>0.30</td>
<td>0.24</td>
<td>0.91</td>
<td>0.00</td>
<td>0.40</td>
<td></td>
</tr>
</tbody>
</table>

Note: Reliability of separation index = 1.00; fixed (all same) chi-square: 2002.0, df: 9; significance: p < .00

Inter-rater reliability was estimated as 0.956 by means of Cronbach’s alpha. Table 1 is a rater-measurement report executed by the FACETS. The average of raw scores was 1.66, and the average of the adjusted scores was 1.64, with a standard deviation of 0.3 and 0.24 respectively. The fourth column shows the severity of the raters expressed in logits, indicating severe ratings with positive values and lenient ones with negative values. The most severe rater was Rater G, with 1.77 logits, and the most lenient rater was Rater C, with -1.97 logits. The discrepancy between the most severe and the most lenient raters was 3.74 logits, which was a considerable degree of difference. The value of standard deviation, 0.91, is also evidence of great variability among the raters. Reliability of the separation index demonstrated 1.00, which was very high, and a chi-square of 2002.0 with 9 df was significant at p < .00. The high reliability index indicated that there were substantial degrees of difference among the raters. The findings demonstrated equivalent or larger significant differences among the raters as compared to the findings of other studies (Lumley and McNamara, 1995; Wigglesworth, 1993; Tyndall and Kenyon, 1996). As the error was as small as 0.07, the rater difference did not seem to depend on error. Fit values, which denote whether or not the data meets the Rasch model, were within the range of the two standard deviations around the mean (0.99 ± (0.40×2)). The studies mentioned above demonstrate that training cannot eliminate rater difference but can increase internal consistency, and so does this result. This suggests that utilizing raw scores may be dangerous. Lunz, Wright and Linacre (1990) assert the necessity of statistical adjustments for rater characteristics to ensure the dependability of scores.

It should be noted that initially there were eleven raters; however, the preliminary analysis indicated that one of the raters was misfitting. As the inconsistent ratings of the misfitting rater cannot be
modelled and compensated for by the FACETS program, the rater was excluded from all subsequent analysis because the rater was considered to be outside the range of two standard deviations and to demonstrate a large number of unexpected responses. Ultimately, the rating data of the ten raters were used for the analysis.

4.2 Rating categories

Table 2 Trait Measurement Report for Six Rating Categories

<table>
<thead>
<tr>
<th>Coherence</th>
<th>Fluency</th>
<th>Interaction</th>
<th>Accuracy</th>
<th>Range</th>
<th>GLOBAL</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
<td>1.7</td>
<td>1.7</td>
<td>1.6</td>
<td>0.02</td>
</tr>
<tr>
<td>1.63</td>
<td>1.64</td>
<td>1.66</td>
<td>1.67</td>
<td>1.69</td>
<td>1.70</td>
<td>1.66</td>
<td>0.02</td>
</tr>
<tr>
<td>0.15</td>
<td>0.08</td>
<td>0.02</td>
<td>-0.01</td>
<td>-0.10</td>
<td>-0.13</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>1.12</td>
<td>1.03</td>
<td>1.16</td>
<td>0.99</td>
<td>0.90</td>
<td>0.86</td>
<td>1.01</td>
<td>0.11</td>
</tr>
<tr>
<td>0.89</td>
<td>0.98</td>
<td>0.84</td>
<td>1.02</td>
<td>1.13</td>
<td>1.16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 is a trait measurement report produced by the FACETS analysis. The multi-faceted Rasch analysis allows us to calibrate different categories all together because it can estimate the step difficulty for each rating scale independently, without averaging over scales. By using this feature, in this study, the CEFR Global Oral Assessment Scale and CEFR Oral Assessment Criteria Grid were calibrated together. The rating categories for analysis were range, accuracy, fluency, interaction, and coherence, and the holistic rating category was GLOBAL. In the CEFR rating scale, Below A1 was replaced by 0, and A1, A2, B1, B2, C1, and C2 were arranged from 1 to 6 respectively, yielding seven levels in total. The “Raw scores” shown in the table are the participants’ scores given by the ten raters, and the “Adjusted scores” indicate the “measures” calibrated with the aid of FACETS.

The average value of raw scores was 1.6, and the average value of adjusted scores or “measures” was 1.66. Considering that the possible maximum score was 6.00, the score on the whole was low. This may be because, firstly, many of the junior high students received ratings Below A1 (=0), and secondly, there were no proficient users who would receive C1/C2, even from among the university students. The participants in the study were not from low-level educational institutions; hence, they did represent the average English speaking ability among all Japanese students.

The fourth column demonstrates the difficulty of the rating categories using degrees of difficulty or “logits”. These indicate high degrees of difficulty with positive values and low degrees with negative values on the same interval scale. In this analysis, the difficulty of the rating categories and the severity of the raters were set at zero in order to obtain the participants’ relative speaking ability. The difficulty measurement report indicates that the most harshly scored category was coherence (0.15 logits) and the
most leniently scored category was GLOBAL (~0.13 logits). The difficulty span between the most leniently and most harshly rated categories was not great (0.28 logits). As the standard deviation was also very small (0.10), these rating categories seemed to have very little difference. Nonetheless, the reliability of a separation index was by no means high (0.69), and the chi-square of 19.5 with 5 df was significant at \( p < .00 \) and, hence the calibration indicated that a significant disparity in difficulty was found among the six rating categories. The fifth column demonstrates that the level of error was small and equal among the categories (0.5), showing evidence of measurement accuracy for this rating scale/grid. The sixth column demonstrates the degree of match, or fit, between the expected model and the observed data. The infit mean square - usually considered the most informative value - for all of the six categories was close to the expected value of 1. Throughout this study, infit mean square values given in the FACETS output within the range of two standard deviations around the mean are regarded as acceptable, based on McNamara (1996) and O'Loughlin (2001). Estimated discrimination on the far right column equally indicates that the categories fit the Rasch model, as the values are within 0.5 to 1.5 (Linacre, 2000).

Table 3 Category Statistics for the Advance of Step Difficulties

<table>
<thead>
<tr>
<th>Category score</th>
<th>Counts used</th>
<th>%</th>
<th>Cum. %</th>
<th>Average measures</th>
<th>Outfit MnSq</th>
<th>Measure</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (BelowA1)</td>
<td>914</td>
<td>11%</td>
<td>11%</td>
<td>-7.53</td>
<td>1.0</td>
<td>-7.41</td>
<td>0.05</td>
</tr>
<tr>
<td>1 (A1)</td>
<td>2653</td>
<td>33%</td>
<td>44%</td>
<td>-4.95</td>
<td>1.0</td>
<td>-3.51</td>
<td>0.04</td>
</tr>
<tr>
<td>2 (A2)</td>
<td>3298</td>
<td>41%</td>
<td>85%</td>
<td>-1.59</td>
<td>1.0</td>
<td>1.18</td>
<td>0.05</td>
</tr>
<tr>
<td>3 (B1)</td>
<td>907</td>
<td>11%</td>
<td>96%</td>
<td>1.23</td>
<td>1.0</td>
<td>3.21</td>
<td>0.07</td>
</tr>
<tr>
<td>4 (B2)</td>
<td>301</td>
<td>4%</td>
<td>100%</td>
<td>2.66</td>
<td>1.2</td>
<td>6.53</td>
<td>0.27</td>
</tr>
<tr>
<td>5 (C1)</td>
<td>15</td>
<td>0%</td>
<td>100%</td>
<td>4.67</td>
<td>0.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5 (C2)</td>
<td>0</td>
<td>0%</td>
<td>100%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3 shows the range of step difficulties, indicating whether the rating scales functioned reliably. The first column indicates the rating scales in which Below A1 was replaced with 0, and A1 to C2 with 1 to 6, respectively. The second to fourth columns show the total number and percentage of ratings given by the ten raters to the participants in each step. The A2 rating, intended for Basic Users, was the most common rating given to the students at 3,298 counts, accounting for 41% of all the ratings, followed by the lower rating of A1 with 2,653 counts, accounting for 33%. These two ratings made up nearly three-quarters of the total. One-tenth of the students were rated as Below A1. Only 15% of the participants were considered as Independent Users (B1/B2). C1, intended for Proficient Users, was given to only 15, accounting for 0.02%, and no students were rated as C2.

The fifth and sixth columns demonstrate fit statistics, in terms of the validity of rating scales. According to the rating scale guidelines (Linacre, 1997), average measures should advance monotonically in order to be good rating scales. The report satisfied this indispensable condition.
Furthermore, the outfit mean square values were around 1, which met the condition to fit the model, namely that the expected outfit mean square is 1 and should not exceed 2. The two columns on the right point to step difficulty calibrations, which need to advance by at least 1.4 but by less than 5.0 logits according to the same guidelines. An examination of the step calibration measures in Table 3 indicates that the step difficulties in this data fell within the acceptable range.

4.3 Participants' speaking ability

The FACETS analysis indicated that the five students proved to be overfit, demonstrating smaller infit mean square values outside the range of two standard deviations around the mean (0.99 ± (0.40×2)), which was not regarded as a major issue. The seven participants demonstrated larger fit statistics, which was considered to be more problematic, representing 5.18% of the test-takers. The number of misfitting candidates was nearly the same as in the result of Lynch and McNamara (1998), indicating a large proportion. Looking carefully into the interactions of the participants, the seven underfit students exhibited mixed characteristics; most were either talkative but made many mistakes, or quiet but with potential speaking ability. The raters were likely to make evaluations on two extremes, that is, either on the participant's fluency/grammatical mistakes for the talkative students, or on the speaker's potential speaking ability/small amount of talk for the less talkative speakers. Therefore, the speakers' talkativeness, which resulted in dominance during the interaction, seems to have become problematic in gaining accurate ratings. It may be crucial for the raters to share their views and opinions about the problem with other raters on the training session.

Figure 1 displays the number of students by different educational institutions per logit. The left side of the figure indicates larger negative values, meaning lower speaking ability, and the right side shows positive values, meaning higher ability. The average speaking ability of the junior high school students was −6.27 logits (the adjusted score was 0.80, which is equivalent to A1 in the CEFR rating), that of the senior high students was −2.40 logits (1.74 and A2 respectively), and that of the university students was 0.06 logits (2.36 and B1 respectively) indicating clear development phenomena by different educational institutions. Examining this in detail, the junior high students remain on the left and do not overlap greatly with the senior high or university students; by contrast, the senior high and university students do overlap to a
considerable degree. Yet, there were no senior high students who obtained ratings of 3 logits or above. Furthermore, the peaks of the number of students between the senior high and university students indicate approximately 2 logits. In addition, these students showed a larger variance as compared to the junior high students.

5. Discussion and Conclusion

A multi-faceted Rasch analysis provided information about the group oral test format in the situation where the participants and the raters were Japanese. With regard to the raters, they showed more leniency on the holistic-rating scale than the other analytical sub-scales, which may suggest the difficulty of employing the holistic-rating only. The findings demonstrated equivalent or larger significant differences among the raters as compared to the findings of other studies. The differences do not stem from errors or internal inconsistency. Calibrating with the aid of multi-faceted Rasch analysis helps to find a rater who shows great variability. The CEFR rating categories and their levels were proved to be valid, as the errors were small and equal among the categories. Nonetheless, the fact that 85% of the participants were categorized as Below A1, A1, and A2, suggests that more detailed criteria for lower levels need to be created for Japanese learners of English. Looking into the characteristics of the participants who displayed underfit values in their speaking ability, the raters had varied opinions about the participants who showed both high and low language abilities (e.g., those displaying fluency with less accuracy). It might be possible to control such variances by building a consensus among raters by having them identify certain attributes in the speakers. There were no Proficient Users (C1, C2) in the study, and the participants’ overall speaking ability was judged as low, which might demonstrate the outcome of the current situation of English education in Japan. Although 46 countries around the world have already employed the CEFR criteria, it may be difficult to introduce the CEFR rating scales in Japan without any modifications.

As mentioned earlier, the group oral is considered to be affected by the interlocutors; indeed, there are more uncontrollable variables in comparison to other test situations. However, these variables may be reduced by utilizing the multi-faceted Rasch analysis. For this reason, this test format can be utilized for performance exercises, assessment, and placement tests in each institution. In addition, the group oral enables beginner level students to interact more easily compared to other types of activities. The participants gave positive feedback on the interaction, for instance, “I enjoyed talking with other people” and “I have never had such an experience before. I would like to do it again.” This kind of activity may be suitable for students in Japan, where lively speaking activities are yet to become popular. Not only European countries but also some Asian countries and regions such as Korea, China, Taiwan, and Hong Kong that regard English as very important have already started utilizing the group oral. This test format may be introduced in Japan before long and more research dealing with the group oral is imperative.
References


