Effects of Test Types and Interlocutors’ Proficiency on Oral Performance Assessment

Junko NEGISHI
Tsurumi University

Abstract

The study aimed to explore the impacts of three types of test and interlocutors’ proficiency level on the test scores of Japanese learners of English participating in a single-speaker task, paired oral discussion, and group oral interaction. The assessment was carried out utilizing the criteria of the CEFR-J (Tono, 2013), the Japanese version of the Common European Framework for Reference (CEFR; Council of Europe, 2001). Five raters assessed 24 university students’ oral performance. The ratings were then analyzed utilizing multi-faceted Rasch measurement (MFRM), and the author confirmed that the raters fit the model for the subsequent main analysis. In the main analysis, first, the MFRM indicated that the speakers scored higher in the group oral interaction compared to the single-speaker task and paired oral discussion. Second, the MFRM analysis showed significant differences across test types and proficiency levels. In addition, the raw score analysis revealed a certain disparity in terms of proficiency levels in that the middle-level speakers received higher scores in the paired and group oral interactions compared to the lower- and upper-level speakers. With regard to the paring and grouping conditions, it was difficult to identify any particular tendency, as various phenomena were seen. Third, the rating criteria showed that the scale was over-segmented for assessing the speakers and invited further investigation.

1. Introduction

In 1976, Folland and Robertson proposed group discussion as a battery of oral performance tests. They described the discussion in comparison with the interview as follows:

The discussion is controlled by the testees, and the examiner is there to evaluate, according to fixed criteria, the linguistic content alone. There is also the advantage that because the examiner does not participate, the testees have more time, and perhaps even inclination, to speak the language. (Folland & Robertson, 1976, p. 161)

In addition to the advantages mentioned above, interaction between examinees in paired or group oral discussion (paired or group oral, hereafter) promotes more communicative, speaking-focused
teaching and learning, which may reflect classroom situations and induce washback, because collaborating during speaking activities helps develop learners’ oral proficiency. Assuming that learning should take place as the result of small group collaboration and cooperation in a classroom, teachers should recreate similar conditions when evaluating students (Webb, 1995). Linking an oral communication task and a classroom activity may suit educational contexts (Davis, 2009). Furthermore, it provides a crucial connection between test scores and a language use situation, which is an essential part of score interpretation (Bachman & Palmer, 1996). Oral interaction also enables speakers to show more varied interaction patterns and language functions within a richer discourse, which leads to authenticity (Taylor, 2000, 2001; Skehan, 2001; ffrench, 2003; He & Dai, 2006). The more authentic the test task becomes, the more valid the score interpretation (Ockey, 2009). Oral interaction, therefore, has various advantages.

2. Background

The paired test format has been implemented with the Cambridge First Certificate in English (FCE) and Certificate of Proficiency in English (CPE) since the 1980s. Group oral discussions are currently used in Asian countries as well. For example, in South Korea, group oral tests have been run by universities for university entrance exams and by the Educational Testing Service (ETS) to select scholarship recipients. Group orals are used in Hong Kong as the Hong Kong Use of English Test and in China as the College English Test – Spoken English Test and the Public English Test Systems. In Japan, however, it is still not common to put oral performance tests into practice for entrance examinations, and paired or group orals are seen far less, even in English proficiency tests. Reasons for not administering paired or group orals might stem from their disadvantages.

In comparison with a single-speaker test, paired/group orals are said to have some deficits. The biggest drawback is that the interaction might be affected by uncontrollable variables from other interlocutors that threaten its validity, such as gender, age, familiarity, status, introvert/extrovert personalities, willingness to speak, and different levels of proficiency. These variables may have a smaller impact on a single-speaker test format such as an interview, monologue, or picture description task. Moreover, compared with an interview, assessments of group oral interactions generally demonstrate lower inter-rater agreement, which is also a threat to validity (Van Moere, 2006).

Among diverse kinds of research, studies related to interlocutor effects have been substantially carried out. One type of research compares the effects of interlocutor personalities (e.g., Berry, 1997, 2004; Van Moere, 2006; Ockey, 2009) or familiarities (Foot, 1999; O’Sullivan, 2002). Another type investigates whether or not an interlocutor’s proficiency level affects scores. Iwashita (1996), for example, found out that high-proficiency interlocutors performed better when paired with other high-proficiency speakers. Low-proficiency speakers also did better when they
interacted with high-proficiency interlocutors compared to low-proficiency interlocutors; however, there was a large variability among individuals. This result agrees with that of Berry (2004); nonetheless, it should be noted that speakers’ increased production of talk when paired with high-proficiency interlocutors did not lead to significantly different scores. Davis (2009) also examined the effect of interlocutor proficiency on the FCE paired test. Second language learners of English were divided into two levels, higher and lower, and they had one conversation with an interlocutor of similar proficiency and another with an interlocutor of different proficiency (i.e., higher or lower). No significant impacts of interlocutor proficiency were observed in the Rasch analysis ability measures, although the lower proficiency examinees produced more talk when matched with higher-level interlocutors, indicating results comparable with those of Berry and Iwashita. However, to my knowledge, there is no study that compares the effects of interlocutor proficiency among the three test types, that is, a single-speaker situation, and paired/group orals. This connects to Research Question 2 described in Section 3.

A smaller number of studies have examined the variability between different test types. Brooks (2009) compared scores in paired and interview tests and found that the examinees were assigned higher scores when they interacted with other interlocutors than when they interacted with interviewers. Shohamy, Reves, and Bejarano (1986) reported that the scores on a group oral correlated with three different one-on-one tasks (an oral interview, picture task, and reporting task); however, the group oral showed the least overlapping variance with the three one-on-one tasks. Fulcher (1996) investigated the variance caused by three different task types. The first task was a one-on-one interview (picture description task) and subsequent group discussion on a related topic, which is similar to the FCE test. The second task resembled the IELTS interview test in which examinees read a text given in advance and answered multiple-choice questions. The third task was a 15-minute discussion preceded by a 10-minute preparation time with a card on which many ideas were written. The questionnaire results showed that the examinees felt least anxious and produced more talk in the discussion task. Fulcher concluded that although the task difference had an impact on scores, the effect was small and negligible. He rather thought that “large task effects may be an artefact of the rating scale used” (1996, p. 37), which relates to Research Question 3. There have been no studies, to my knowledge, that compared the impacts of the three test types (i.e., a single-speaker test, paired oral, and group oral), which will be addressed as Research Question 1.

3. Research Questions

The study aimed to explore the following three research questions (RQs) using data from Japanese learners of English who took three different oral tests and were assessed by raters.

RQ1: Are there effects on raters’ scores on the three different oral performance tests: a single-speaker test, paired oral test, and group oral test?
RQ2: Are there effects on raters’ scores when the participants are paired or grouped with interlocutors with different proficiency levels? If so, are these effects present for participants of all proficiency levels?

RQ3: Do the rating criteria function reliably?

The investigation of RQ1 may contribute to determining whether paired or group oral test types can elicit scores similar to those of a single-speaker test format designed to measure the communicative ability of Japanese learners of English. In addition, the answer to RQ2 will tell us whether or not deliberate pairing/grouping is required at speakers’ proficiency level when implementing paired/group orals in Japan in the future. Also, while exploring RQ1 and RQ2, we should know whether the rating criteria used in the study function reliably to assess the speakers (RQ3; see Section 4.3).

4. Method

4.1 Participants

The participants were 24 Japanese university students who had been studying English for at least 6 years. They took the three oral performance tests as part of their classroom activities. The 12 students from university A were taking a cross-cultural, online distance learning course as an elective class with students in another country. The class included presentations and discussions. No students from university A majored in English; however, their overall English speaking ability was relatively high, especially as some were returnees from overseas. The other 12 students, from university B, were all English majors, and one student had experience studying abroad. The participants were recruited from these two universities in order to assemble a sample that included a wide range of English speaking abilities. The author explained the research project to the participants, and they signed written consent forms. The project was approved by the Ethical Review Board at the author’s university (#1121).

<table>
<thead>
<tr>
<th>Gender</th>
<th>English Ability (TOEIC® Score)</th>
<th>Former English Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>University A</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>University B</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

The three tests were administered at each university so that the students were paired/grouped with interlocutors whom they already knew. This was done to avoid possible assessment variability, since O’Sullivan (2002) and Foot (1999) reported the effects of familiarity on scores. Table 1 shows the basic data of the participants. The male to female ratio was 11 to 13,
and the range of TOEIC® scores was approximately 600 to 960 at university A and 350 to TOEFL® 90 at university B.

4.2 Test Types and Procedures

The following are the three tests the participants completed. The participants’ oral performances were video recorded:

Test A—A single-speaker oral production task (picture description): A four-panel cartoon of STEP Pre-1st test, with Eiken permission, was used. The participants had one minute to think about the story, looking at the cartoon, and narrate the story within two minutes. All the students completed Test A first.

Test B—Paired oral interaction task: The author, who was the participants’ teacher as well, paired them according to their proficiency levels assessed by Test A. For the first round, they were paired with interlocutors who had similar proficiency levels and talked freely about a prompt, “family.” For the second round, they were paired with interlocutors of a different proficiency level and talked about “school.” They interacted for about 4 minutes, and the talk was later cut to 200 seconds when a DVD was burned for the subsequent assessment and analysis.

Test C—Group oral interaction task: The author grouped the students by the same procedure employed in Test B. For the first round, three students who had similar proficiency levels discussed a prompt, “dreams.” For the second round, one or two student(s) who had different proficiency levels were grouped and discussed “English.” The interaction lasted more than 5 minutes, and the talk was later cut to 300 seconds.

First, Test A was given to all the students, and Tests B and C were partially counterbalanced because of the class schedule and the students’ attendance conditions.

4.3 Rating Criteria and Assessment

Five Japanese raters, who had been teaching English for more than 10 years and held at least an M.A., assessed the participants. The author did not participate in the assessment. The CEFR-J (ver. 1.1; Tono, 2013), the Japanese version of the Common European Framework of Reference, (CEFR; Council of Europe, 2001) was utilized as the rating criteria.

The original CEFR was designed not only for assessment but also for providing a learning goal for language learners. It is intended for all the purposes involved in teaching and creating syllabi, curriculum guidelines, examinations, and textbooks. The CEFR includes the Common Reference Levels providing its basic framework as A1 and A2 for Basic Users, B1 and B2 for Independent Users, and C1 and C2 for Proficient Users. The framework has been introduced not

---

1 According to Van Moere (2006), who reported that the topic or prompt was not a significant factor, this study employed four different topics.
only within the multi-linguistic and multi-cultural sphere of the European Union but also in other countries including Japan and other Asian countries.

Tono and many linguists produced the Japanese version of the CEFR and named it CEFR-J (Tono, 2013). Like the CEFR, the CEFR-J contains “CAN-DO” descriptors that tell learners what their language learning goals are. The Ministry of Education, Culture, Sports, Science and Technology (MEXT) has invited junior high schools and high schools to put the “CAN-DO list” into practical use in order to set up language learning goals (2013). The biggest difference between the CEFR and the CEFR-J is that the latter criteria are more segmented in order to fit the proficiency levels of Japanese learners of English. This is based on the fact that about 80% of Japanese learners of English are classified as level A (Negishi, 2011; Negishi, Takada, & Tono, 2012). The levels of the CEFR-J are as follows: Pre-A1; A1.1, A1.2, A1.3, A2.1, and A2.2 for level A; B1.1, B1.2, B2.1, and B2.2 for level B; and C1 and C2 for level C. The original CEFR has qualitative aspects of spoken language use, i.e., range, accuracy, fluency, interaction, and coherence, while the CEFR-J does not. For this reason, the assessment was carried out holistically in the present study.

The rating criteria used for the study were “speaking: presentation” for Test A and “speaking: interaction” for Tests B and C. An example of the monologue descriptors for A2.2 is “Can give a short speech using a series of simple words and sentences directly related to the speaker’s everyday life (about him/herself, schools, or local matters),” and an example of the interactional descriptors for A2.2 is “Can exchange his/her opinions and thoughts, deliver his/her pros and cons, and compare things and people” (originally written in Japanese; the author’s translation).²

Four out of the five raters had already received training and had previously assessed other participants using the original CEFR criteria. The five raters, including a new rater, first trained together by watching the CEFR training video (North & Hughes, 2003). Next, they carried out a trial assessment while watching a similar video and had a vital discussion utilizing the CEFR-J criteria before conducting the actual assessment.

4.4 Multi-Faceted Rasch Measurement (MFRM) Analysis

In performance-based assessment, a speaker gives a performance and a rater assesses the speaker utilizing a rating scale. Interactions between the speaker and the test instrument or between the rater and the speaker are seen. For this reason, a more complex procedure is, in turn, required to estimate the speaker’s performance, underlying abilities, or skills. An analysis utilizing a multi-faceted Rasch measurement (MFRM) is often used to estimate the speaker’s performance under this condition (McNamara, 1996). In the present study, the MFRM was utilized as it

² The CEFR-J is downloadable from http://www.cefr-j.org/
allowed us to transform raw scores into interval scales called *measures*. In order to conduct the MFRM, the software program FACETS 3.7.1.4 was used (Linacre, 2014).

The data were defined as having three facets: the severity of the raters, the difficulty of the three types of test, and the ability of the participants. The MFRM executed through the FACETS program produced each of the individual facets in separate tables. First, to what extent the raters consistently assessed the speakers will be investigated before the main analyses.

5. Results and Discussion

5.1 Rater Measurement Report by the MFRM

The raters assessed the participants by watching the DVDs for Tests A, B, and C from Pre-1 to C2 of the CEFR-J; however, no participants were assessed as Pre-1 or C2. For the subsequent analysis, the ratings were transformed into numbers: 1 for A1.1, 2 for A1.2, ... 9 for B2.2, and 10 for C1.

Table 2.

<table>
<thead>
<tr>
<th>Rater Measurement Report for the Five Raters</th>
<th>Observed Average</th>
<th>Fair Average</th>
<th>Measure (Logits)</th>
<th>Model SE</th>
<th>Outfit MeanSq</th>
<th>Estimated Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>5.22</td>
<td>5.18</td>
<td>0.15</td>
<td>0.08</td>
<td>0.87</td>
<td>1.24</td>
</tr>
<tr>
<td>Rater 2</td>
<td>5.61</td>
<td>5.81</td>
<td>-0.12</td>
<td>0.08</td>
<td>1.12</td>
<td>0.89</td>
</tr>
<tr>
<td>Rater 3</td>
<td>5.57</td>
<td>5.75</td>
<td>-0.09</td>
<td>0.08</td>
<td>0.90</td>
<td>1.03</td>
</tr>
<tr>
<td>Rater 4</td>
<td>5.84</td>
<td>6.17</td>
<td>-0.28</td>
<td>0.08</td>
<td>0.96</td>
<td>1.11</td>
</tr>
<tr>
<td>Rater 5</td>
<td>4.96</td>
<td>4.76</td>
<td>0.34</td>
<td>0.08</td>
<td>1.10</td>
<td>0.91</td>
</tr>
<tr>
<td>Mean</td>
<td>5.44</td>
<td>5.54</td>
<td>0.00</td>
<td>0.08</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.35</td>
<td>0.56</td>
<td>0.25</td>
<td>0.00</td>
<td>0.12</td>
<td></td>
</tr>
</tbody>
</table>

*Note: Fixed (all same) chi-square: 40.7; df: 4; significance: p = .00*

Table 2 is a five-rater measurement report executed by FACETS. The mean average of the speakers’ averaged raw scores was 5.44, and the mean value of the fair average (adjusted scores) calibrated with the aid of the FACETS was 5.54, with a standard deviation of 0.35 and 0.56, respectively. The fourth column, “Measure”, 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 5, with 0.34 logits, and the most lenient rater was Rater 4, with -0.28 logits. The discrepancy between the most severe and most lenient raters was 0.62 logits, which is not a

---

3 Logits are the values expressed by units in the measurement scale called “log odds units.” The odds are expressed as a logarithm (“log” for short) of the naturally occurring constant e. The logit scale is an interval scale: “It can tell us not only that one item is more difficult than another, but also how much more difficult it is” (McNamara, 1996, p. 165). In other words, the distance between a certain interval (e.g., between -2 and -3 and between 1 and 2) may not be the same in raw scores but is the same when expressed in logits.
considerable degree of difference. The value of the standard deviation, 0.25, represents moderate variability among the raters. The model standard error (SE) was 0.08, which is small. However, the chi-square of 40.7 with 4 df was significant at \( p = .00 \). Hence, the null hypothesis that all raters were equally severe must be rejected. The fit values, outfit mean square in the sixth column, which denote whether or not the data meet the Rasch model, were within the range of two standard deviations around the mean \((0.99 \pm 0.12 \times 2)\), which is regarded as acceptable, based on McNamara (1996) and O’Loughlin (2001). In other words, there were no raters who were overfit (i.e., too consistent) or underfit (i.e., too inconsistent). Therefore, the raters were considered self-consistent. The estimated discrimination in the far right column equally indicated that the raters fit the Rasch model, as the values were within the range of 0.5 to 1.5 (Linacre, 2000). Although the raters were not equally severe/lentient, the rater report allows us to apply the assessment data for the subsequent analysis.

5.2 Differences Among the Three Tests (RQ1)

5.2.1 Observed Data

Table 3.

Descriptive Statistics for the Three Types of Test

<table>
<thead>
<tr>
<th>Test</th>
<th>N</th>
<th>Total</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>120</td>
<td>658</td>
<td>5.48</td>
<td>2.58</td>
<td>0.187</td>
<td>-1.171</td>
</tr>
<tr>
<td>B</td>
<td>240</td>
<td>1266</td>
<td>5.28</td>
<td>2.54</td>
<td>0.012</td>
<td>-1.212</td>
</tr>
<tr>
<td>C</td>
<td>240</td>
<td>1341</td>
<td>5.59</td>
<td>2.48</td>
<td>-0.249</td>
<td>-1.150</td>
</tr>
</tbody>
</table>

Note. N = 120 for Test A means 24 participants \( \times \) ratings by 5 raters; N = 240 for Tests B and C means 24 participants had paired/group orals twice \( \times \) ratings by 5 raters.

Figure 1. Histograms of the five raters’ cumulative raw scores for Test A, Test B, and Test C

First, the raw scores assigned by the five raters based on the CEFR-J rating scales are reported here. The descriptive statistics for the three tests are in Table 3. All the participants took Tests B and C twice (paired/grouped interaction with similar/different level interlocutors), and the numbers of raters’ cumulative scores in Tests B and C were approximately double the number in
Test A. As the mean scores indicate, Test B yielded the lowest score, 5.28; Test C, the highest, 5.59; and Test A, the middle score, 5.48. The standard deviations were 2.58, 2.54, and 2.48 for Tests A, B, and C, respectively. These are quite large deviations. Considering that the maximum possible score was 11 (C2), the score on the whole was approximately in the center. In terms of skewness, Test A was slightly right skewed with a value of 0.187, and Test C was left skewed to some degree with a value of -0.249. Negative kurtosis suggests platy distributions.

Figure 1 presents histograms of the five raters’ cumulative raw scores for the three types of test. No histograms display normal distributions because the cumulative ratings in the middle level were fewer than expected. In comparison with Test A, there were greater numbers of higher scores for B1.2 in Tests B and C, and for B2.1 in Test C. The data were not normally distributed, and the MFRM was an effective means to normalize the given data.

5.2.2 Three Test Format Report by the MFRM

Table 4.
Measurement Report for the Three Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Observed Average</th>
<th>Fair Average</th>
<th>Measure (Logits)</th>
<th>Model SE</th>
<th>Outfit MeanSq</th>
<th>Estimated Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test C</td>
<td>5.59</td>
<td>5.78</td>
<td>-0.10</td>
<td>0.05</td>
<td>0.73</td>
<td>1.26</td>
</tr>
<tr>
<td>Test A</td>
<td>5.48</td>
<td>5.61</td>
<td>-0.02</td>
<td>0.08</td>
<td>1.45</td>
<td>0.62</td>
</tr>
<tr>
<td>Test B</td>
<td>5.28</td>
<td>5.27</td>
<td>0.12</td>
<td>0.05</td>
<td>1.02</td>
<td>1.02</td>
</tr>
<tr>
<td>Mean</td>
<td>5.44</td>
<td>5.54</td>
<td>0.00</td>
<td>0.06</td>
<td>1.06</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.16</td>
<td>0.26</td>
<td>0.11</td>
<td>0.01</td>
<td>0.36</td>
<td></td>
</tr>
</tbody>
</table>

*Note: Fixed (all same) chi-square: 8.4; df: 2; significance: p = .01*

Table 4 is a test format measurement report for the three tests produced by the MFRM. The observed average shows the participants’ averaged raw scores, and the fair average shows the adjusted scores calibrated by the FACETS program. The fourth column demonstrates the difficulty of the test types in logits. The measure with a positive value indicates that Test B was the most difficult, while those with negative values mean easier tests; that is, the most leniently scored test was Test C, and Test A was in the middle. As the span of difficulty between the most leniently and most harshly rated tests was not great (0.22 logits) and the SE values were also very small (0.05 and 0.08), these tests seem to have very little difference. However, the chi-square of 8.4 with 2 df was significant at p = .01, which is evidence for rejecting the null hypothesis that all test types are equally difficult. The sixth column shows the degree of match, or fit, between the observed and the expected data. The outfit mean square values given in the FACETS output with the range of two standard deviations around the mean (1.06 ± 0.36 × 2) are regarded as acceptable. Estimated discrimination points out that the test types fit the Rasch model, as the values are within 0.5 to 1.5.
As a result, the answer to RQ1 is that statistically significant difference was confirmed between the three tests; that is, there were effects on raters’ scores on the three different oral performance tests.

5.3 Test Types and Proficiency Levels (RQ2)

5.3.1 Observed Data

First, the analysis using the observed data (raw scores) is reported in this section for the reason that the MFRM is not capable of measuring each participant’s ability in different test format but is capable of measuring each participant’s overall ability in all the test formats.

![Figure 2. Differences of the raw scores between Test B and Test A](image)

![Figure 3. Differences of the raw scores between Test C and Test A](image)

Figure 2 shows the differences between Tests B and A, and Figure 3 those between Tests C and A when the participants were divided into three proficiency levels, namely, higher-level, middle-level, and lower-level, with eight participants at each level. In Figure 2, the bar graphs display the differences of average raw scores between Tests B and A when the participants were matched with a similar- or different-level interlocutor. By the same token, Figure 3 exhibits the differences of average raw scores when the Test A score was subtracted from the scores of Test C when the participants were placed with a similar- or different-level interlocutor. The bar graph with a negative value (facing downwards) expresses that the speakers received lower scores compared with the Test A situation, and the bar graph with a positive value (facing upwards) indicates that the performances obtained higher scores in comparison with those of Test A.

As can be seen, the higher-level participants tended to receive lower scores in the Test B and C conditions irrespective of pairing or grouping patterns. Although the lower-level participants showed a similar phenomenon, the disparity was slightly smaller. In contrast to the higher- and lower-level participants, the middle-level speakers (whose TOEIC® scores were approximately 500 to 700) earned higher scores on Tests B and C than on Test A. This result
might tell us that the speakers at the middle level can exhibit their speaking ability better when paired or grouped than when speaking alone. By contrast, the higher-level speakers are likely to be unable to demonstrate their true ability in the paired or grouped condition. As the higher-level speakers earned lower scores when matched with different level speakers, that is, with the middle- or lower-level speakers, they may have played down their speaking ability in order to comfort their interlocutors or they may have given their floor to their interlocutors. To consider that the order of tests is a cause of the difference would be difficult because, if so, the middle-level speakers should exhibit the similar phenomenon. The lower-level speakers may have performed worse in paired or group orals because they had stronger anxiety about being asked to speak with their interlocutors.

In terms of pairing or grouping, the higher-level speakers showed a larger disparity between Test A and Test B/C when they were matched with different proficiency level interlocutors than when they were matched with similar level interlocutors. On the other hand, the lower-level speakers tended to display a larger difference in scores when similar proficiency level interlocutors were placed in the discussion. The middle-level speakers did not exhibit any specific tendencies in terms of pairing or grouping. It is difficult to identify any particular tendency as various phenomena were seen in the pairings and groupings.

5.3.2 Test Types and Proficiency Measurement Report by the MFRM

Table 5.

Five Test Format Measurement Report

<table>
<thead>
<tr>
<th></th>
<th>Observed Average</th>
<th>Fair Average</th>
<th>Measure (Logits)</th>
<th>Model SE</th>
<th>Outfit MeanSq</th>
<th>Estimated Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test C (similar)</td>
<td>5.61</td>
<td>5.81</td>
<td>-0.12</td>
<td>0.08</td>
<td>0.59</td>
<td>1.41</td>
</tr>
<tr>
<td>Test C (different)</td>
<td>5.57</td>
<td>5.75</td>
<td>-0.09</td>
<td>0.08</td>
<td>0.87</td>
<td>1.10</td>
</tr>
<tr>
<td>Test A</td>
<td>5.48</td>
<td>5.61</td>
<td>-0.03</td>
<td>0.08</td>
<td>1.45</td>
<td>0.62</td>
</tr>
<tr>
<td>Test B (different)</td>
<td>5.39</td>
<td>5.46</td>
<td>0.04</td>
<td>0.08</td>
<td>1.16</td>
<td>0.87</td>
</tr>
<tr>
<td>Test B (similar)</td>
<td>5.16</td>
<td>5.08</td>
<td>0.20</td>
<td>0.08</td>
<td>0.87</td>
<td>1.18</td>
</tr>
<tr>
<td>Mean</td>
<td>5.44</td>
<td>5.54</td>
<td>0.00</td>
<td>0.08</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.18</td>
<td>0.29</td>
<td>0.13</td>
<td>0.00</td>
<td>0.33</td>
<td></td>
</tr>
</tbody>
</table>

Note: Fixed (all same) chi-square: 10.7; df: 4; significance: p = .03

Table 5 is a test format measurement report produced by the MFRM in order to confirm the differences of test types and proficiency levels statistically. The observed average is the raw scores given by the five raters, and the fair average indicates the measures calibrated with the aid of FACETS. The mean value of the observed average was 5.44, and the mean value of the fair average or measures was 5.54. The fourth column shows the test format difficulty utilizing degrees of difficulty or logits. Positive measure values indicate that the test format is more difficult, while negative values represent an easier format; as can be seen, Test C was the easiest
when similar-level students were grouped (−0.12 logits), resulting in higher scores. The measurement report also indicates that the most harshly scored test format was Test B (0.20 logits) when the participants were paired with a similar level interlocutor. The difficulty span between the most leniently and most harshly rated test format was not great (0.32 logits). As the standard deviation was also small (0.13 logits), these test formats, including proficiency levels, seem to have little difference. The SE values were also small (0.08 logits). Nevertheless, the chi-square of 10.7 with 4 df was significant at p = .03; therefore, the null hypothesis that all tests are equally difficult is rejected; that is, the calibration indicated that a significant disparity in test format was found among the five test formats. The sixth column indicates whether or not the test format fits the Rasch model in terms of the size and direction of the difference between expected and observed data for each test. Throughout this measurement, the outfit mean square values given in the FACETS output were within the range of two standard deviations around the mean (0.99 ± 0.33 × 2) and were regarded as acceptable. The estimated discrimination with values of 0.5 to 1.5 also proves that the test formats fit the Rasch model.

The investigation of the difficulty of test formats with the aid of the MFRM revealed a statistical difference between the test formats and participants’ proficiency levels. It should be noted that the analysis using raw scores also exhibited a certain amount of disparity caused by the three types of test and proficiency differences. Thus, the answer to RQ2 is that there were observable effects and statistical differences on raters’ scores when the participants took Tests A, B, and C and were paired/grouped with interlocutors with similar/different proficiency levels.

### 5.4 Rating Scale Statistics (RQ3)

#### Table 6.

**Rating Scale Statistics**

<table>
<thead>
<tr>
<th>Category Score</th>
<th>Data</th>
<th>Quality Control</th>
<th>Rasch-Andrich Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count used</td>
<td>%</td>
<td>Cum. %</td>
</tr>
<tr>
<td>1 (A1.1)</td>
<td>28</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>2 (A1.2)</td>
<td>63</td>
<td>11%</td>
<td>15%</td>
</tr>
<tr>
<td>3 (A1.3)</td>
<td>84</td>
<td>14%</td>
<td>29%</td>
</tr>
<tr>
<td>4 (A2.1)</td>
<td>72</td>
<td>12%</td>
<td>41%</td>
</tr>
<tr>
<td>5 (A2.2)</td>
<td>41</td>
<td>7%</td>
<td>48%</td>
</tr>
<tr>
<td>6 (B1.1)</td>
<td>65</td>
<td>11%</td>
<td>59%</td>
</tr>
<tr>
<td>7 (B1.2)</td>
<td>93</td>
<td>16%</td>
<td>74%</td>
</tr>
<tr>
<td>8 (B2.1)</td>
<td>79</td>
<td>13%</td>
<td>88%</td>
</tr>
<tr>
<td>9 (B2.2)</td>
<td>57</td>
<td>10%</td>
<td>97%</td>
</tr>
<tr>
<td>10 (C1)</td>
<td>18</td>
<td>3%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Table 6 shows the range of step difficulties, indicating whether the rating scales function reliably. The first column shows the rating scales. The second to fourth columns show the total number and percentage of ratings given by the five raters to the participants in each step. The B1.2 rating was the most common rating given to the speakers at 93 counts, accounting for 16% of all the ratings. The two extremes, A1.1 (5%) and C1 (3%), accounted for very few ratings, while other ratings had comparable frequencies, from 10% to 16% except for A2.2 (7%). A possible reason for the platy distribution is that there were not many participants (N = 24). Also, they were recruited to have a wide range of English speaking abilities (TOEIC® 350 to 960) and to be scattered in the range, which resulted in the platy distribution. Even so, the small number of A2.2 ratings is puzzling. The fifth and sixth columns demonstrate fit statistics, in terms of the validity of rating scales. According to the rating scale guidelines (Linacre, 1997; Smith, Wakely, de Kruif, & Swartz, 2003), average measures should advance monotonically in order for the rating scales to be good. The report satisfies this indispensable condition. The expected outfit mean square values were from 0.5–1.5, meeting the condition to fit the model. 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. The average step calibration measures were 0.78, which indicates that the step difficulties in the data are too small, except for between B2.2 and C1 (3.33 – 1.71 = 1.60 logits). Specifically, the advancement from A2.2 to B1.1 was −0.49, a reverse advancement, which is likely the cause of the puzzling phenomenon.

What can be drawn from the analysis in order to answer RQ3? It seems that the CEFR-J rating scale was over-segmented for the middle-level speakers. A more detailed investigation is required.

6. Conclusion

The study presented three research questions about the impacts on scores caused by the test types and proficiency levels, and about the rating scale. First, the assessment was carried out utilizing the CEFR-J rating criteria. The raters’ ratings were then analyzed, and the author confirmed that the raters fit the Rasch model for the subsequent main analysis. Basically, the analyses were carried out with the aid of MFRM as the assessed data were not normally distributed. The results were as follows:

In terms of RQ1, “Are there effects on raters’ scores on the three different oral performance tests: a single-speaker test (Test A), a paired oral test (Test B), and a group oral test (Test C)?”, Tests C seemed to enable the participants to achieve higher scores compared to Test A. Test B was likely to be the most severe for the participants among the three tests. The aim of presenting RQ1 was to obtain information on whether Test B or C could produce similar scores to Test A. As there was a statistically significant result, it might be recommended to conduct Test B or C as a battery of varied performance tests rather than to replace Test A with Test B or C.
In relation to RQ2, "Are there effects on raters' scores when the participants are paired or grouped with interlocutors with different proficiency levels?", the MFRM indicated significant differences by proficiency levels; furthermore, a certain disparity was observed with regard to the raw scores. Specifically, the middle-level speakers demonstrated a tendency to earn higher scores on Tests B and C than on Test A. Grouping or pairing conditions, in terms of the participants' proficiency level, were not likely to have a specific effect on their scores.

The last research question, "Do the rating criteria function reliably?", will lead to further research. The answer to RQ3 was confirmed by analyzing step difficulties in the MFRM in that each step was too small. The rating criteria employed in the study, the CEFR-J, basically demonstrate learning goals by means of CAN-DO lists for English learners, and this did not seem to fit the aim of assessing the speakers of the study. Utilizing the CAN-DO list as a learning goal is effective since each step is set little by little while the step is over-segmented for assessing speakers of this type.

A limitation of the study is that the number of participants was not sufficiently large. The sample size of 24 participants was selected in consideration of the raters' quality and work burden. If there had been a greater number of participants, the results might have been generalizable. Another limitation was that the investigation was solely quantitative. Future research should use a qualitative method to explore the discourse and interaction patterns of the three tests. Moreover, future research can be carried out to produce more adequate rating criteria for assessing multiple speakers, including investigations about the criteria's validity and reliability.

Acknowledgement

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant Number 25884073.

References


MEXT. (2013). Kaku chu., koto-gakko no gaikokugo kyoiku ni okeru “CAN-DO list” no katachi deno gakushu totatsu no mokuhyo settei no tame no tsubiki [A guidebook for setting learning goals using a “CAN-DO list” in foreign language teaching at junior high schools and high schools]. Tokyo: Taishukan.


