A Multifaceted Rasch Analysis of Rater Reliability of the Speaking Section of the GTEC CBT

Rie KOIZUMI  
*Juntendo University*  
Yasuko OKABE  
*Center for Entrance Examination Standardization*  
Yuko KASHIMADA  
*Benesse Corporation*

**Abstract**

Second language (L2) speaking assessment can be affected by raters as well as tasks and other factors. High-stakes speaking tests require that high rater reliability be assured and that such information be reported to the public. In Japan, investigations into rater reliability and the use of multifaceted Rasch analysis have been limited for L2 speaking assessment in both high-stakes contexts and classroom situations. To fill this gap, this study examines the rater reliability of the Speaking Section of the Global Test of English Communication Computer Based Testing (GTEC CBT). This test has nine tasks for evaluation and 23 assessment criteria. We analyzed 648 test takers’ responses using multifaceted Rasch analysis. The results showed that raters differed in severity to a small degree but demonstrated high rater agreement and rater self-consistency. The bias analysis indicated a small percentage of systematic biased patterns between raters and test takers and 25.78% of biases between raters and criteria. Implications for improving assessment were discussed.

**1. Introduction**

The Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan announced that for second language (L2) English university entrance examinations in the near future, externally made tests of speaking and writing will be used along with a nationally made test of reading and listening and that externally made tests of four skills will eventually be used (MEXT, 2016). While there are several candidates for externally made tests that can be used for this purpose (Eigo yonginou shikaku kentei shiken kondankai, 2016), each test has its own characteristics in terms of test content and delivery/administration methods and how the test institution (or test developer/provider) publicizes its information on test quality. All test providers have the responsibility of communicating their test content and quality to the public so that test
stakeholders, particularly test users, including university teachers and high school students and teachers, can interpret test scores and use them appropriately for their decision making (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Introducing the assessment of L2 speaking ability into university entrance examinations seems crucial in terms of maintaining relevance across teaching contents, tasks that learners can encounter in real life, and what is assessed in a test (i.e., test construct) and in terms of the potentially strong impact of speaking tests on teaching and learning (see Akiyama, 2003). However, the introduction of speaking tests may be useful only when test users can interpret and use test scores adequately. Therefore, it is necessary for test providers to conduct test validation and publicize the results, especially considering the high-stakes nature of university entrance examinations (Japan Language Testing Association [JLTA] Steering Committee & JLTA Language Testing Terminology Committee, 2006).

Validity includes various aspects that need to be examined in multiple stages, and validation continues recursively from the test planning phase to the phase of test analysis based on operational test data (see for details, Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2008). Messick (1996) presented six aspects of validity that needed to be examined: (a) content aspect (i.e., whether the test content is relevant to and representative of the test construct or what should be assessed), (b) substantive aspect (whether the process of taking the test is consistent with the process that the test developer intended to elicit), (c) structural aspect (whether the structure within the test accords with the one that the test developer intended to have), (d) generalizability aspect (whether test scores are similar or generalizable across test administration, test forms, test items that assess similar ability, or raters, or whether reliability is high), (e) external aspect (whether the test scores are related to scores from other tests or external criteria derived from real-life performances, in a way and to a degree predicted by a theory), and (f) consequential aspect (whether the test use is appropriate).

While investigations into and reporting of all six aspects are usually essential in high-stakes testing, reliability is a basic aspect that must be examined at an early stage of validation (e.g., Chapelle et al., 2008). In particular, speaking assessment involves rater judgment and requires firm evidence of high reliability for high-stakes decisions. Rater reliability is low if raters consistently fail to produce similar scores when rating the same speeches and if they fail to maintain the same rating quality throughout the rating session. Periodic investigations and reporting are needed to assess whether adequate procedures are set and practiced. Although some may argue that rater reliability is assured after rigorous rater training, it is often reported that even after intensive training, raters may continue to have idiosyncratic tendencies, such as (a) rating too strictly or too leniently, and (b) avoiding extreme scores and giving middle scores (Eckes, 2011). Consequently, examining rater reliability is important for maintaining test quality. The current study examines and reports the rater reliability of the Speaking Section of the Global Test of
Studies on rater reliability of L2 speaking tests have been generally limited in Japan. For example, searching the CiNii database (Scholarly and Academic Information Navigator organized by the National Institute of Informatics) for papers on rater reliability using the term “rater reliability” and “speaking” (or “oral”) returned a limited number of papers: one from ARELE (Aso, 2000), three from university bulletins, and two conference abstracts. They all addressed classroom-based speaking tests. For example, Aso (2000) conducted oral interviews and asked 10 teacher raters to evaluate 10 high school students’ utterances using both holistic and analytic rating scales. He examined the reliability among raters and within the same raters and found high rater reliability in both holistic and analytic ratings. Such analysis will be of use for developing and modifying speaking tests and test procedures and for enhancing language assessment literacy of teachers and test stakeholders.

Additionally, previous studies on rater reliability basically analyze raw scores (or raw ratings) using correlations, Cronbach’s alpha, t-tests, or other basic analytical methods. Now, a more useful method is readily available called multifaceted Rasch analysis, but its use has been limited in Japan despite its popularity internationally (e.g., Eckes, 2005; Hagan, Pill, & Zhang, 2016). There are some exceptions for which multifaceted Rasch analysis is used for speaking tests in classrooms (Akiyama, 2003; Fukazawa, 2010; Negishi, 2015; Koizumi, In’nami, & Fukazawa, 2016). For instance, in Akiyama (2003), 219 junior high school students took two of the four tasks, and five teacher raters evaluated their talk using the evaluation criteria of fluency, vocabulary, grammar, intelligibility, and overall task fulfillment. The data were examined using multifaceted Rasch analysis, and differences in task difficulty and evaluation criteria and fits of test-takers, tasks, and criteria to the data were reported, but rater reliability was not reported. In another example, Negishi (2015) used three speaking formats—a monologue format, a paired oral format, and a group oral format—with 24 university students. Five raters evaluated a talk, and ratings were analyzed using multifaceted Rasch analysis. It was found that raters differed in severity or leniency and that raters evaluated consistently throughout the rating session, but Negishi did not report information on some aspects of rater reliability, that is, rater agreement and systematic biased patterns among raters, test takers, and formats. More extensive and thorough use of multifaceted Rasch results would lead to more insights into rater behaviors and areas for improvement.

Multifaceted Rasch analysis is an extension of basic Rasch analysis that analyzes two facets (e.g., test takers and tasks). Rasch measurement is used to conduct Rasch analysis. It broadly belongs to the family of item response theory (IRT) and is a measurement model for converting ordinal test scores to interval scores on the logit scale (Bond & Fox, 2015). Multifaceted Rasch measurement is useful in analyzing test data affected by three or more facets, such as test takers, tasks, raters, and evaluation criteria. It calibrates multiple facets on the same scale to enable users to compare the facets on the same frame of reference.
In terms of rater reliability, multifaceted Rasch analysis is advantageous over analysis based on raw scores in the framework of classical test theory, in the following four main points (Eckes, 2011). First, multifaceted Rasch measurement can produce in-depth information, such as information on rater severity, rater self-consistency (or within-rater consistency), and systematic, unexpected, biased patterns (called rater bias) in relation to test takers, tasks, evaluation criteria, or other facets (Barkaoui, 2013; Eckes, 2011; McNamara & Knoch, 2012). These rater results can be readily compared to those of other facets in the same frame of reference. It is true that some results can be obtained through analysis of raw scores. For instance, researchers can examine rater severity by conducting a t-test or an analysis of variance (ANOVA) to investigate if there are differences between scores from two or three raters. However, researchers cannot derive other results such as rater self-consistency and rater bias. Another example is rater agreement. Although the percentage of agreement can be computed using raw scores, multifaceted Rasch analysis also gives the percentage at which raters are expected to agree, so researchers can examine if an actual agreement is higher than the expected agreement. Multiple perspectives of rater reliability can help detect abnormalities of rater behaviors.

Second, details can be derived through multifaceted Rasch analysis from just a single run of analysis, and the procedures are basically simple. Although detailed, careful analysis of raw scores may suggest important patterns to be noticed, it requires more time and expertise to reach such interpretations than multifaceted Rasch analysis.

Third, multifaceted Rasch analysis can deal with data with missing responses. It is usually difficult to ask two or more raters to evaluate responses of all tasks from all test takers, and it is common to ask different raters to rate responses of different tasks from different test takers. When analyzing rater reliability using raw scores, researchers have to choose between a few options for coping with missing data—such as considering different raters as the same rater, imputing values for missing responses, or even deleting such data. While there may be situations in which such treatments are needed, they may not produce accurate results. In contrast, multifaceted Rasch analysis solves such problems as long as responses are substantially connected with each other (see Eckes, 2011 for details).

Fourth, multifaceted Rasch measurement can provide fair scores, that is, scores adjusted by considering differences in rater severity and criterion difficulty and interpreted on the same scale as raw scores (Eckes, 2011). Thus, these scores are less affected by construct-irrelevant factors. All in all, it can be concluded that more studies would benefit from the application of multifaceted Rasch analysis.

2. Current Study

The current study examines rater reliability of the GTEC CBT Speaking Section using multifaceted Rasch analysis. The GTEC CBT Speaking Section was selected because rater
reliability has not been reported despite the section’s high-stakes nature. We investigated the following research questions (RQs).

RQ1: Do raters differ in severity?  
RQ2: Do raters agree highly with each other?  
RQ3: Do raters perform consistently throughout the rating session?  
RQ4: Do raters show systematic biased patterns against evaluation criteria or test takers?

The analysis of and report on rater reliability of a high-stakes test will be useful for test users to grasp the characteristics of the speaking test. It will also serve as an example of detailed analysis of rater reliability using multifaceted Rasch analysis, which can be employed readily in other contexts of speaking or writing assessment to improve assessment practices.

3. Method

3.1 Test Takers and Raters

This study used the part of 2015 data from the GTEC CBT Speaking Section with 648 test takers and 13 certified raters. Data were derived from the actual operational test, which was held on one of three occasions during the year, and was administered and evaluated based on normally used, regulated procedures.

The test takers consisted of first- to third-year senior high school students (70%) and teachers and those who seemed to be interested in taking the test. The tasks and criteria of the whole Speaking Section were included in the analysis.

Raters had experiences of rating the GTEC CBT Speaking Section for two years or more. All raters majored in English at an undergraduate level, six had experiences in teaching L2 English, and some had experiences in scoring other tests. To be certified, all raters need to show high standards of rating, as described in section 3.3. Out of 13 raters who evaluated for the current data set, 6 had more experience in rating speaking performances and served as the third raters (see section 3.3 below).

3.2 Test Contents and Procedures of the GTEC CBT Speaking Section

According to Benesse Corporation (2017), the GTEC CBT was developed by the Benesse Corporation and Center for Entrance Examination Standardization. It is intended to assess English proficiency in four skills for academic purposes, and is meant to be used primarily for Japanese and overseas university entrance examinations. The Speaking Section lasts approximately 20 minutes and consists of tasks presented on a computer screen. Test tasks reflect situations where university students speak English monologically and dialogically in class or in daily life. The Speaking Section has three parts: Listening and responding (6 tasks; e.g., “Where do people like
to travel to in your country?”; no planning time and 10- to 20-second response time); Delivering and asking for information (3 tasks; e.g., describing and comparing dance centers; asking questions to staff at a dance center; 30-second planning time and 60-second response time); Expressing your opinion (3 tasks; e.g., “Do you think technology has changed the way we live?”; 0- to 90-second planning time and 1- to 3-minute response time). Each part is intended to be equal to A2 to B2 levels of the Common European Framework of Reference (CEFR; Council of Europe, 2001). Each task in a part has one to five evaluation criteria (i.e., rubrics), and one criterion has one to three points that count as a full mark (see Table 1). Since each task has a different number of criteria focusing on different aspects of speaking ability, criteria, not tasks, were used as one facet of the analysis.

While the test was administered, responses to each task from each test taker were automatically audio-recorded and transmitted to the test center. Raters listened to each response uploaded onto a computer and entered their rating in the computer.

Table 1

Evaluation Criteria of the GTEC CBT Speaking Section

<table>
<thead>
<tr>
<th>Part</th>
<th>Task</th>
<th>Criteria (Full mark)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 Answer a conversational question (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 Answer a conversational question (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 Answer a conversational question (1)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1 a. Provide factual information (1), b. Provide factual information (1), c. Provide factual information (1), d. Provide factual information (1), e. Include an appropriate introduction and closing (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 a. Give a preference (1), b. Provide two reasons (2), c. Include an appropriate introduction and closing (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 a. Ask two questions (2), b. Include the reason for calling and a closing phrase (2)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1 a. Provide an opinion (1), b. Provide one reason that supports the opinion (2), c. Provide one reason that supports the opinion (2), d. Fluency (3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 a. Present a position (1), b. Provide a relevant reason that supports the position (2), c. Provide a relevant reason that supports the position (2), d. Fluency (3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 a. Answer a follow-up question (2), b. Hesitation (2)</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Rater Training and Monitoring

Rater candidates attended a one-week rater training session, in which they evaluated sample answers and responses using evaluation criteria under the supervision of two raters with more experience. After the training, rater candidates took a certifying test, to allow the selection of only those candidates who had demonstrated sufficiently accurate rating performance. After they were
certified as raters, they participated in a rating session of the field test. They received feedback from experienced raters to understand the quality standard. Then, two trial rating sessions followed, using actual responses from a field test. In both sessions, when two raters gave a different score for the same response, an experienced rater evaluated it again. Raters who gave a score different from the experienced rater received feedback and a chance to modify their rating style. If they received the same feedback in the first and second sessions, they were not allowed to participate in the operational rating session, which was used for this study. Raters who participated in the operational rating session were also required to undertake English training, including listening to lectures and responding to quizzes, to maintain their English proficiency.

For each task, a speech was evaluated independently by two raters, using one or more criteria. When raters gave divergent ratings, a third, experienced rater evaluated the speech. The experience raters determined the final rating with the knowledge that this utterance received different scores from the first two raters. The experience raters were allowed to refer to the first and second ratings. This rating procedure was the same as the one usually implemented in the operational test.

3.4 Analysis

We analyzed rating data and used Facets (Version 3.71.4; Linacre, 2014) to employ the partial credit model of multifaceted Rasch measurement, since we expected each criterion to have different structures. We included three facets, that is, test takers, criteria, and raters, and centered the criterion and rater facets with test takers floated on the logit scale. We also included the data from the first, second, and third raters for one task because this allowed us to examine the overall picture of the GTEC CBT Speaking Section. It should be noted that the results were similar when we did not include the data of the third rater.

Regarding RQ1, rater measures on the logit scale were used to examine rater severity. RQ2 used the percentage of agreement between raters. RQ3 was investigated using Infit and Outfit mean square statistics of the rater facet. RQ4 employed bias analysis (interaction analysis) between the three facets.

Following Eckes (2005), two criteria for judging mean squares were used: A wide range of a criterion had mean squares of 0.50 to 1.50, which were considered to fit to the Rasch model (Linacre, 2013). Those of less than 0.50 were considered to be overfit, whereas those of more than 1.50 were considered underfit. However, those of less than 0.50 and of between 1.50 and 2.00 were considered to be acceptable. Those of more than 2.00 were considered to have negative impact on the measurement. In contrast, a narrow, strict range of a criterion had mean squares of 0.70 to 1.30. When values are out of a certain range, this means that elements of a facet (e.g., test takers) do not conform to the expectation of the Rasch model. Overfit means that there are fewer variations and that response patterns are too predictable. In contrast, underfit means that there are more variations than the Rasch model can predict and that response patterns are too unpredictable.
Both Infit and Outfit mean-square values were used with two ranges of criteria, since the GTEC CBT is a high-stakes test and requires rigorous examination.

4. Results and Discussion

The Facets output showed that the variance explained by Rasch measures was 48.06%, which was more than 20% and sufficient for the data to be considered unidimensional (Engelhard, 2013, p. 185). We had only 2 responses (0.005%, out of 37,014 valid responses) with absolute standardized residuals equal to or greater than |±2.00| and 0 responses with such residuals equal to or greater than |±3.00|. According to Eckes (2011), 5% or less of such residuals in case of |±2.00| and 1% or less in case of |±3.00| are appropriate, and we concluded that the data fit sufficiently to the Rasch model overall.

Figure 1 represents distributions of three facets (Test takers, Criteria, and Raters) on the logit scale, with a mean of 0 and positive values on the scale indicating test takers with higher speaking ability, more difficult criteria, and more severe raters. Test takers and criteria were distributed widely in terms of proficiency and difficulty, respectively, and raters varied to some extent. Table 2 shows a summary of the three facets. The test taker facet had strata of 6.35, which indicates that the test could stratify test takers into six different ability groups. Since the GTEC CBT Speaking Section is intended to cover three CEFR levels, this result was promising.

Table 3 shows the results of fit statistics to the Rasch model. There were 8.33% to 22.84% of underfitting test takers (depending on what mean-square ranges and which type of fit statistics were used). Underfitting test takers could display a few patterns, but one pattern is that they achieved high scores in difficult criteria and can be considered to have high speaking ability, but they obtained low scores in easy criteria. The relatively high percentage of underfitting test takers is explored below in the bias analysis. The criteria facet had strata of 29.84, meaning that criteria had about 30 levels of difficulty, which is appropriate for a high-stakes test. Outfit mean squares of criteria showed underfitting criteria with unexpected rated patterns (21.74%), which need to be examined further.

RQ1 asks about differences in rater severity. The measures of rater severity in Table 2 show that the severity ranged from −0.41 to 0.17 logits (0.58 logit difference). Rater severity levels were found to be significantly different ($\chi^2 = 110.5, df = 12, p < .00$). Furthermore, the rater strata of 3.64 and high rater reliability of .86 also suggest that raters differed in severity. However, the difference in rater severity was equivalent to 0.24 raw-score points at most (in fair scores, 1.35–1.11, which is not shown in Table 2), and the range of rater severity on the logit scale was limited compared to the one of test-takers’ speaking ability (i.e., about one-twentieth; 0.048 = 0.58/12.04). Thus, differences in rater severity were considered to be small.

As mentioned above, multifaceted Rasch analysis can produce test-takers’ ability scores called fair scores (i.e., scores with differences in rater severity and criterion difficulty adjusted;
Both Infit and Outfit mean-square values were used with two ranges of criteria, since the GTEC CBT is a high-stakes test and requires rigorous examination.

4. Results and Discussion

The Facets output showed that the variance explained by Rasch measures was 48.06%, which was more than 20% and sufficient for the data to be considered unidimensional (Engelhard, 2013, p. 185). We had only 2 responses (0.005%, out of 37,014 valid responses) with absolute standardized residuals equal to or greater than $|\pm 2.00|$ and 0 responses with such residuals equal to or greater than $|\pm 3.00|$. According to Eckes (2011), 5% or less of such residuals in case of $|\pm 2.00|$ and 1% or less in case of $|\pm 3.00|$ are appropriate, and we concluded that the data fit sufficiently to the Rasch model overall.

Figure 1 represents distributions of three facets (Test takers, Criteria, and Raters) on the logit scale, with a mean of 0 and positive values on the scale indicating test takers with higher speaking ability, more difficult criteria, and more severe raters. Test takers and criteria were distributed widely in terms of proficiency and difficulty, respectively, and raters varied to some extent. Table 2 shows a summary of the three facets. The test taker facet had strata of 6.35, which indicates that the test could stratify test takers into six different ability groups. Since the GTEC CBT Speaking Section is intended to cover three CEFR levels, this result was promising.

Table 3 shows the results of fit statistics to the Rasch model. There were 8.33% to 22.84% of underfitting test takers (depending on what mean-square ranges and which type of fit statistics were used). Underfitting test takers could display a few patterns, but one pattern is that they achieved high scores in difficult criteria and can be considered to have high speaking ability, but they obtained low scores in easy criteria. The relatively high percentage of underfitting test takers is explored below in the bias analysis. The criteria facet had strata of 29.84, meaning that criteria had about 30 levels of difficulty, which is appropriate for a high-stakes test. Outfit mean squares of criteria showed underfitting criteria with unexpected rated patterns (21.74%), which need to be examined further.

RQ1 asks about differences in rater severity. The measures of rater severity in Table 2 show that the severity ranged from $-0.41$ to $0.17$ logits ($0.58$ logit difference). Rater severity levels were found to be significantly different ($\chi^2 = 110.5, df = 12, p < .00$). Furthermore, the rater strata of 3.64 and high rater reliability of $0.86$ also suggest that raters differed in severity. However, the difference in rater severity was equivalent to $0.24$ raw-score points at most (in fair scores, $1.35 - 1.11$, which is not shown in Table 2), and the range of rater severity on the logit scale was limited compared to the one of test takers' speaking ability (i.e., about one-twentieth; $0.048 = 0.58/12.04$). Thus, differences in rater severity were considered to be small.

As mentioned above, multifaceted Rasch analysis can produce test takers' ability scores called fair scores (i.e., scores with differences in rater severity and criterion difficulty adjusted; Figure 1. Variable map of multifaceted Rasch analysis. Information on criterion category threshold measures was omitted due to space limitations.)
Table 2
Descriptive Statistics of Three Facets

<table>
<thead>
<tr>
<th></th>
<th>M (SD)</th>
<th>Min to Max</th>
<th>Range</th>
<th>Strata</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test takers</td>
<td>0.92 (1.44)</td>
<td>-6.94 to 5.10</td>
<td>12.04</td>
<td>6.35</td>
<td>.95</td>
</tr>
<tr>
<td>Criteria</td>
<td>0.00 (1.54)</td>
<td>-3.18 to 3.35</td>
<td>6.53</td>
<td>29.84</td>
<td>1.00</td>
</tr>
<tr>
<td>Raters</td>
<td>0.00 (0.17)</td>
<td>-0.41 to 0.17</td>
<td>0.58</td>
<td>3.64</td>
<td>.86</td>
</tr>
<tr>
<td>Third raters only</td>
<td>0.02 (0.08)</td>
<td>-0.11 to 0.11</td>
<td>0.22</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 3
Percentages of Mean-Square Fit Statistics

<table>
<thead>
<tr>
<th></th>
<th>value &lt; 0.70 (overfit)</th>
<th>0.70 ≤ value ≤ 1.30 (fit)</th>
<th>1.30 &lt; value &lt; 1.50 (overfit)</th>
<th>value &lt; 0.50 (underfit)</th>
<th>0.50 ≤ value ≤ 1.50 (overfit)</th>
<th>value &lt; 1.50 (underfit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test takers</td>
<td>16.20</td>
<td>66.98</td>
<td>16.36</td>
<td>2.16</td>
<td>89.04</td>
<td>8.33</td>
</tr>
<tr>
<td>Criteria</td>
<td>32.36</td>
<td>44.14</td>
<td>22.84</td>
<td>9.88</td>
<td>74.07</td>
<td>15.59</td>
</tr>
<tr>
<td>Raters</td>
<td>4.35</td>
<td>73.91</td>
<td>21.74</td>
<td>0.00</td>
<td>100.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note. There were three test takers who had perfect scores and did not get fit values. aThere were 0.93% with Infit mean squares of more than 2.00, and 6.02% with Outfit mean squares of more than 2.00. bThere were 0.00% with Outfit mean squares of more than 2.00.

Eckes, 2011). As long as multifaceted Rasch analysis is used with a rater facet included, test users should be free from worries about rater severity differences. However, at the level of raw scores, differences in rater severity need to be addressed. Even though the third, experienced rater makes a final decision in the GTEC CBT Speaking Section when the first two raters disagree (see section 3.3), there must be cases where a speech is scored by two lenient or two severe raters. In that case, the first two raters agree on a rating, and this will be a final score. Thus, the test-takers’ speaking ability could be overestimated or underestimated. Therefore, more training and monitoring will be needed for raters with relatively severe and lenient raters.

Additionally, the third raters need to be more accurate than other raters and they should be neither severe nor lenient, since the third, experienced raters evaluate the speech for the final rating when the first and second raters disagree in rating. Regarding the results of such raters (n = 6), the mean of rater severity measures was close to zero (M = 0.02) and standard deviation was also small (SD = 0.08). However, two raters had severity measures of –0.11 and 0.11, and they may need to adjust their ratings to more precise ratings.

RQ2 explores if there was high agreement between raters. The percentage of agreement...
between raters was 79.4%, which was higher than 63.4%, which is the percentage that Rasch analysis predicts that there should be this level of agreement after considering differences in rater severity. Thus, rater agreement, or reliability between raters, was sufficiently high.

However, rater agreement of 79.4% means that in 20.6% of cases, raters did not reach agreement. We further explored in what criterion there were more disagreements. Table 4 shows 10 criteria that had low agreement percentages. The commonality was that all had a full mark of 2 or 3 and required higher degrees of judgment. Particularly, out of six criteria with less than 70% agreement, five criteria were about whether a test taker provided one or two reasons (e.g., Part 2 Task 2 [b] Provide two reasons; see Table 1). Judgments of whether a test taker provides reasons are related to the prompt and what he/she said before and may differ across raters. If this is the case, this specific information can be included in the rater training. For example, more sample talks should be provided in training criteria that are more likely to diverge than other criteria with more agreement.

Table 4
Agreement Percentages of Less Than 80% for Each Criterion

<table>
<thead>
<tr>
<th>Score difference</th>
<th>Part 2 Task 2</th>
<th>Part 2 Task 3</th>
<th>Part 3 Task 1</th>
<th>Part 3 Task 1 (b)</th>
<th>Part 3 Task 3</th>
<th>Part 3 Task 2 (b)</th>
<th>Part 3 Task 2 (c)</th>
<th>Part 3 Task 2 (d)</th>
<th>Part 3 Task 3 (a)</th>
<th>Part 3 Task 3 (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>−2</td>
<td>1.08</td>
<td>0.77</td>
<td>0.77</td>
<td>1.08</td>
<td>0.00</td>
<td>1.23</td>
<td>4.01</td>
<td>0.00</td>
<td>0.77</td>
<td>0.31</td>
</tr>
<tr>
<td>−1</td>
<td>13.43</td>
<td>10.34</td>
<td>16.36</td>
<td>14.04</td>
<td>6.79</td>
<td>14.35</td>
<td>15.74</td>
<td>9.10</td>
<td>18.06</td>
<td>14.81</td>
</tr>
<tr>
<td>0</td>
<td>61.57</td>
<td>75.62</td>
<td>64.20</td>
<td>65.74</td>
<td>71.14</td>
<td>63.89</td>
<td>62.50</td>
<td>70.68</td>
<td>69.60</td>
<td>71.45</td>
</tr>
<tr>
<td>1</td>
<td>22.84</td>
<td>11.11</td>
<td>17.90</td>
<td>16.51</td>
<td>20.83</td>
<td>20.06</td>
<td>15.28</td>
<td>19.44</td>
<td>11.42</td>
<td>13.43</td>
</tr>
<tr>
<td>2</td>
<td>1.08</td>
<td>2.16</td>
<td>0.77</td>
<td>2.62</td>
<td>1.23</td>
<td>0.46</td>
<td>2.47</td>
<td>0.77</td>
<td>0.15</td>
<td>0.00</td>
</tr>
</tbody>
</table>

RQ3 investigates self-consistency within raters using two ranges of Infit and Outfit mean squares as criteria (see Table 3). The results show that all raters were within this standard (Infit: \( M = 1.09, SD = 0.09, Min = 0.85, Max = 1.21 \); Outfit: \( M = 1.10, SD = 0.15, Min = 0.89, Max = 1.27 \)) and that the reliability within raters was high. This means that raters tended to provide self-consistent ratings throughout the rating. This includes the third raters who fit to the expectation of the Rasch model even with a narrow range of criterion (Infit: 0.85 to 1.01; Outfit: 0.89 to 1.20).

RQ4 explored systematic, unexpected biased patterns of raters using bias analysis using \( z \) scores (or \( t \) values in the current Facets output; Linacre, 2013, p. 212) of equal to or more than \( \pm 2.00 \). Table 5 showed that there were 2.91% of bias patterns between raters and test takers, which means that some raters evaluated certain test takers more strictly or more leniently than expected. Furthermore, there were 25.78% of bias patterns between raters and evaluation criteria, which means that some raters evaluated certain criteria more strictly or more leniently. There were
4.05% of bias interactions between test takers and criteria, which suggests that some test takers were evaluated more strictly or leniently when certain criteria were used. While 25.78% of bias patterns between raters and evaluation criteria may appear surprising, the high percentage may be because the combinations are limited in number (n = 256) compared to other combinations. Additionally, previous studies reported high percentages of rater x criterion interaction patterns. For example, Eckes (2005) reported 37.0% in the speaking section of the L2 German test. Other studies attempted to reduce high percentages of rater x criterion interaction patterns by innovative methods. One example is Elder, Knoch, Barkhuizen, and von Randow (2005), which examined how biases between raters and criteria can be reduced by giving feedback to raters. These previous studies seem to suggest that rater x criterion bias interaction may be an area that requires special attention in rater training and monitoring.

Table 5
Summary of Bias Analysis

<table>
<thead>
<tr>
<th></th>
<th>Rater x Test taker</th>
<th>Rater x Criterion</th>
<th>Test taker x Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>N combinations</td>
<td>6,351</td>
<td>256</td>
<td>14,835</td>
</tr>
<tr>
<td>% of large z scores(^a)</td>
<td>2.91%</td>
<td>25.78%</td>
<td>4.05%</td>
</tr>
<tr>
<td>Maximum z</td>
<td>3.06</td>
<td>8.77</td>
<td>4.45</td>
</tr>
<tr>
<td>Minimum z</td>
<td>−3.48</td>
<td>−5.24</td>
<td>−5.41</td>
</tr>
<tr>
<td>M</td>
<td>−0.03</td>
<td>−0.01</td>
<td>−0.06</td>
</tr>
<tr>
<td>SD</td>
<td>0.87</td>
<td>1.93</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Note. \(^a\)absolute z scores (standardized bias scores) equal to or greater than |± 2.00|.

Close analysis of the utterances of these cases suggest that some biased patterns between raters and test takers could be explained by deviations from the assigned topic. For example, Part 2, Tasks 1 and 2 require test takers to leave a message on their friend’s answering machine about dance centers, and Part 2, Task 3 requires test takers to leave a message asking questions on an answering machine of a staff member at a dance center. Some students did not seem to notice a situation change regarding whom they talk to, gave a talk to a friend even in Task 3, and obtained lower ratings. Although these responses were unexpected, talks that deviated from the topic deserve lower scores, and this does not seem to be a problem of this test. However, the results of having a small percentage of biases (2.91%) found between certain students and certain raters may indicate that some test takers are less sensitive to their deviations from the topic than others and that some raters are more sensitive to detecting deviations from topics than others and that there are interactions between such test takers and raters. For such test takers, it may be possible to add a warning regarding the change in the situation. Moreover, for raters, the information that there are some test takers giving off-topic responses can be provided in rater training by providing some sample utterances so that raters are aware of and appropriately evaluate such responses.
Additionally, the tendency of obtaining lower scores by providing off-topic responses seems also related to a relatively high percentage of underfitting test takers (e.g., 8.33% in case of using Infit mean squares of more than 1.50), which we observed in Table 3. For systematic biased patterns between raters and criteria and between test takers and criteria, close analysis did not lead to any noticeable overall patterns of biases. In any case, systematic unexpected patterns identified by bias analysis should be used to give feedback to raters for future sophistication of rating quality control and to test developers for future refinement of the test.

5. Conclusion

We examined rater reliability of the GTEC CBT Speaking Section and generally found favorable results. While the rater severity differed across raters to a small degree (RQ1), rater agreement and rater self-consistency were high (RQ2 and RQ3). The bias analysis (RQ4) showed a small percentage (2.91%) of biased patterns between raters and test takers and 25.78% having biased patterns between raters and criteria. Although differences in rater severity were small (0.24 raw-score points), differences in rater severity and biased patterns could lead to different decisions in university entrance examinations; more rater training and monitoring may be needed to maintain a high standard. As summarized in the Introduction, most of these results could not have been obtained without the use of multifaceted Rasch measurement. The current results using the data on one test occasion may not be generalized to all occasions, but may be generalized to some extent because procedures of test administration and scoring are maintained across occasions.

Implications for L2 speaking assessment are as follows: First, the current study can be used as a resource providing positive evidence of rater reliability of the GTEC CBT Speaking Section. Second, this study can serve as an example for those interested in applying fine-grained analysis of rater reliability using multifaceted Rasch measurement for their assessment context. Third, some useful information was derived through the analysis of rater reliability. For example, the results of criteria in which the first and second raters tended to disagree more enable us to predict the criteria that should require more time and resources in rater training. Another example is from the results of bias analysis, which provide cases where some raters seem to detect off-topic responses of some test takers. These results may suggest the need to add some extra emphasis on instructions of the tasks or prompts themselves as well as the need to bring more attention to how to evaluate off-topic responses in rater training sessions. These hints for improving the practices may be hidden in the data but can be revealed through careful analysis mainly using multifaceted Rasch measurement.

Future studies should examine aspects other than rater reliability that the current study did not examine, such as correlation between the GTEC CBT Speaking Section scores and real-life criterion, such as how test takers handle English in academic contexts. Messick’s (1996) six aspects of validity to be examined, which were presented in the Introduction, may serve as a
guiding tool. Investigations into more aspects of validity would lead to more positive or negative evidence for validity of GTEC CBT test scores, or to be more specific, validity of interpretation and use based on GTEC CBT test scores (since validity is not inherent in a test or test scores, but the degree of validity varies depending on how test scores are interpreted and used; Messick, 1996). Furthermore, while the current analysis was conducted once, regular analyses using multifaceted Rasch measurement should be conducted as quality assurance to maintain high rater reliability and give constant feedback to improve rating performance. Feedback based on results from multifaceted Rasch analysis, such as on rater severity, rater agreement, rater self-consistency, and rater bias, may help raters improve their rating performance and eventually enhance rater reliability of the test (see Elder et al., 2005, for an example of feedback to raters based on Facets results).

Acknowledgments

This study was partially funded by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C), Grant number 26370737. We would like to thank three anonymous reviewers and Yo In’nami for constructive comments on an earlier draft.

References


Appendix: An excerpt of the input (specification) file for the Facets analysis (with bias analysis of criteria and raters) Note. This information is necessary when replicating the current analysis. B = bias analysis. See Koizumi (2016) for running Facets.

title = GTEC_CBT_speaking
convergence = 0.1 ; size of largest remaining marginal score residual at convergence
unexpected = 2 ; size of smallest standardized residual to report
uncenter = 1 ; examinee facet floats
pos = 1 ; examinees greater score greater measure
pt-biserial = Yes ; report the point-biserial correlation
noncenter = 1 ; examinee facet floats
arrange = m ; arrange output tables in Num descending and Logit ascending order
facets = 3 ; 3 facets
Person, 2 Criteria, 3 Rater
noncenter = 1 ; examinee facet floats
negative = 1 ; for examinees, greater score greater measure
Pt-biserial = Yes ; report the point-biserial correlation
Inter-rater = 3 ; facet 3 is the rater facet
Missing = N
Model=
?.1B,?B,R1
?.2B,?B,R1
?.3B,?B,R1
?.4B,?B,R1
?.5B,?B,R1
?.6B,?B,R1
?.7B,?B,R1
?.8B,?B,R2
?.9B,?B,R1
?.10B,?B,R2
?.11B,?B,R2
?.12B,?B,R2
?.13B,?B,R2
?.14B,?B,R1
?.15B,?B,R2
?.16B,?B,R2
?.17B,?B,R3
?.18B,?B,R1
?.19B,?B,R2
?.20B,?B,R2
; Continues to the right column
?.21B,?B,R3
?.22B,?B,R2
?.23B,?B,R2
?.B1B,?B,R1
?.B23B,?B,R2
?.B1,?B,R1
?.B23,?B,R2
*
Labels=
1, Test takers
1 = … (Test taker identification number comes here.)
648 = (Test taker identification number)
*
2, Criteria
1 = … (Criteria label comes here.)
23 = … (Criteria label)
*
3, Raters
1 = … (Rater identification number comes here.)
13 = … (Rater identification number)
*
data =
1 1 3 1 (Test taker 1, Criteria 1, Rater 3, Rating 1)
648 23 7 1