Papers in Language Testing and Assessment, 2019, Volume 8, Issue 2, 1-33

Examination of CEFR-J spoken interaction tasks using many-facet Rasch measurement and generalizability theory

Rie Koizumi[1], Emiko Kaneko[2], Eric Setoguchi[3],
Yo In'nami[4] & Naoyuki Naganuma[5]

[1]Juntendo University, Chiba, Japan
[2]University of Aizu, Fukushima, Japan
[3]University of California, Los Angeles, USA
[4]Chuo University, Tokyo, Japan
[5]Tokai University, Kanagawa, Japan

Attempts are underway to develop prototype tasks, based on a Japanese version of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001; CEFR-J; Negishi, Takada, & Tono, 2013). As part of this larger project, the current paper reports on the creation of spoken interaction tasks for five levels (Pre-A1, A1.1, A1.2, A1.3, and A2.1). Tasks were undertaken by 66 Japanese university students. Two raters evaluated their interactions using a three-level holistic rating scale, and 20% of the performances were double rated. The spoken ratings were analysed using many-facet Rasch measurement (MFRM) and generalizability theory (G-theory). MFRM showed that all the tasks fit the Rasch model well, the scale functioned satisfactorily, and the difficulty of the tasks generally concurred with CEFR-J levels. Results from G-theory that employed the p x t design, including tasks as a facet, showed the different proportion of variance accounted for by tasks, as well as the number of tasks that could be required to ensure sufficiently high reliability. The MFRM and G-theory results effectively revealed areas for improving spoken interaction tasks; the results also showed the usefulness of combining the two methods for task development and revision.

Key words: CEFR-J, generalizability theory, many-facet Rasch measurement, spoken interaction

Email address for correspondence: rkoizumi@juntendo.ac.jp

Introduction

Recent reforms in education have increasingly focused on relating international frameworks or standards to curriculum, instruction, and assessment at both national and classroom levels. These shifts have occurred to allow for examination and enhancement of educational quality (Papageorgiou, 2016). The Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) has been predominantly employed for such purposes (Papageorgiou, 2016). The adoption of the CEFR has helped to set adequate learning objectives, increasing understanding and collaboration within and across educational contexts (Council of Europe, 2001). Yet, the use of the CEFR has received criticism. For example, usage in such varied contexts has fostered the misunderstanding that when tests are linked with the CEFR, scores from these different tests are strictly comparable. This misunderstanding and consequent misuse have been demystified by recent publications (Deygers, Van Gorp, & Demeester, 2018; Harsch, 2018), which emphasise that each test has its own constructs and purposes and therefore produces results that are not always comparable.

The CEFR has been adopted in Japan primarily due to its relevance and usefulness. According to Shiina (2013), language situations in Japan are similar to those in Europe to some degree, in terms of the heightened need to interact with a variety of people from different countries, to understand diverse cultures and languages, and to maintain better communication between individuals. Schmidt, Runnels, and Nagai (2017) summarise how the CEFR has been introduced into the Japanese context, explaining that after the CEFR (Council of Europe, 2004) was translated into Japanese, its concepts gained popularity gradually and they have since been incorporated into primary, secondary, and tertiary education, for example, in the form of Can-Do descriptors, six levels in the common language framework, or language portfolios. Through this process, it has become clear that there is a need to modify the CEFR to suit the Japanese context, wherein a majority of Japanese learners of English are at basic-user levels (A1 and A2) (Negishi, Takada, & Tono, 2013). This need has led to the creation of the adapted Japanese version of the CEFR, termed CEFR-J (Negishi et al., 2013). The CEFR-J project has the general goal of developing a common language framework suitable for the Japanese context. The purposes of this framework are to promote transparency in discussions of language learning, teaching, and assessment in Japan, and to improve English language teaching in Japan (Negishi et al., 2013). The CEFR-J's level classification is more fine-grained than the CEFR's, especially at lower levels. There are 12 levels overall (Pre-A1, A1.1, A1.2, A1.3, A2.1, A2.2, B1.1, B1.2, B2.1, B2.2, C1, and C2). The CEFR-J covers

the areas of reading, listening, writing, spoken interaction, and spoken production.

According to Tono (2019), one of the current CEFR-J project objectives is to develop CEFR-J-based prototype assessment tasks for each level, and to create a manual on how to adequately construct such tasks. The tasks and manual will be made public, so that teachers and test designers can use them in their particular assessment context (either low-stakes or high-stakes). Additionally, they will be able to refer to prototype tasks and the manual when they develop new CEFR-J-based tasks. As part of a larger project, this article reports on the creation and analysis of such tasks, which are intended to assess second language (L2) spoken interaction ability by using a role-play format conducted by an examiner.

In assessing spoken interaction, the central construct is interactional competence, which refers to "the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event" (Galaczi & Taylor, 2018, p. 226). In this assessment context, a test taker is usually required to interact with an examiner or one or more other test takers. Each type of interaction (i.e., talking with either an examiner or other test takers) has its own weaknesses and strengths (Galaczi & ffrench, 2011; O'Sullivan & Green, 2011). For example, because they are on an equal footing, test takers may talk more naturally with each other and may use more diverse language functions (e.g., starting a new topic) than they would when speaking with an examiner. On the other hand, test takers may not exhibit a high-quality speaking performance in peer-to-peer talk due to the influence of other test takers and their own traits. Moreover, learners with low-level ability may also need some assistance or scaffolding to get involved in pair or group interactions. These weak aspects of peer-to-peer talk can be addressed by examiner-learner talk, especially when examiners follow predetermined procedures. Additionally, even within examiner-learner talk, task types affect spoken interaction, and role-play tasks have been known to enable test takers to produce interactions that are more similar to everyday life conversation than non-scripted interview tasks (Kormos, 1999). These considerations have led us to select a role-play format using a test taker and an examiner, which allows us to cater to beginning levels as well as to relatively advanced levels.

The current study has two purposes: (a) to examine the measurement quality of CEFR-J spoken interaction tasks aligned with the CEFR-J levels Pre-A1 to A2.1 and (b) to examine the degree to which such tasks are aligned with the CEFR-J levels. In addition to the usual procedures for examining measurement quality, we are interested in how many tasks are necessary to assure high reliability in situations

where CEFR-J tasks are used; the fewer the number of tasks required may suggest a higher task precision, which can be an aspect of measurement quality. We used many-facet Rasch measurement (MFRM) and generalizability theory (G-theory) for purpose (a). The two methods are detailed in the next section. As for (b), while spoken interaction tasks have been developed based on the CEFR (e.g., Cambridge English exams; see University of Cambridge ESOL Examinations, 2011), it is not clear to what extent tasks are aligned with more fine-grained CEFR-J levels. This question is investigated using MFRM results. The current results will help identify areas for improvement while moving closer to the goal of developing tasks that can provide useful information on test-takers' spoken interaction ability and their CEFR-J levels.

MFRM and G-theory

MFRM and G-theory are both measurement theories, and have been used in language assessment for multiple decades. MFRM is an extension of the basic Rasch model, which analyses data in terms of factors (called "facets") of both test takers and test tasks. MFRM can include other factors, such as raters. It can provide not only estimates of test-taker ability, task difficulty, and rater severity, and each factor's reliability. It is also able to generate statistics on how each test taker, task, and rater fit the model (i.e., fit statistics), how precisely they are measured (standard errors), and how a rating scale functions (rating scale functioning analysis). Finally, this method can also reveal the degree to which a combination of factors provides unexpected patterns (see e.g., Barkaoui, 2014; Bond & Fox, 2015; Eckes, 2015, 2019; Engelhard & Wind, 2018; McNamara, Knoch, & Fan, 2019).

G-theory, an extension of classical test theory, examines how multiple factors (such as test takers, tasks, and raters) contribute to test score variance. In G-theory, factors other than test takers are termed "facets," while the test-taker factor is named "object of measurement." G-theory can decompose overall test variance into variance components attributable to a number of factors such as test takers, tasks, and raters, as well as variance components attributable to interactions between factors, such as a test-taker-by-task interaction, and a task-by-rater interaction. Through this analysis, we can derive percentages of contribution of each factor and its interaction. Furthermore, depending on whether test scores are to be used for a norm-referenced (or relative) decision, or for a criterion-referenced (or absolute) decision, we can obtain two types of reliability estimates (generalizability coefficient and phi coefficient). We can further calculate reliability estimates when the number of tasks and raters varies (see e.g., Brennan, 2001; Gebril, 2013; Sawaki & Xi, 2019; Webb, Shavelson, & Haertel, 2006).

The combined use of MFRM and G-theory has been encouraged, because each has distinct characteristics, providing complementary information. According to McNamara et al. (2019), they differ on four main points. First, MFRM converts raw scores into measures on an interval logit scale, while separating factors (e.g., test taker, task) in such a way that they are comparable on the same scale. G-theory separates total variance into variance components attributable to different factors and their interactions, examining their impact on the total variance. Second, MFRM offers detailed results at the individual level. MFRM provides estimates of each individual in each factor (e.g., a measure of each test taker and each task). It also reveals measures of precision and results of fit statistics at the individual level, as well as at the group level (such as person reliability and task reliability). In contrast, G-theory provides results only at the group level (such as the degree of impact of factors and their interactions, and overall reliability). Third, MFRM adjusts observed scores, offering "fair average" scores by considering diversity in measures of each factor. For example, when test takers are evaluated by lenient raters, their scores tend to be inflated. "Fair average" scores are adjusted to be lower than observed scores in MFRM, simulating a situation where they are evaluated by raters with average leniency. In contrast, G-theory does not adjust scores. Fourth, MFRM can treat missing data flexibly. G-theory can handle missing data, with limited flexibility. By using MFRM and G-theory in combination, test developers and users can obtain rich information at both individual and group levels for the purposes of test development, evaluation, and research (McNamara et al., 2019). See Eckes (2015) and Engelhard and Wind (2018), for further comparisons between MFRM and G-theory.

Despite encouragement for pairing the two methods, such a pairing is not commonplace. Wind and Peterson (2018) summarised the most dominantly employed methods in previous first language (L1) and L2 assessment studies, specifically those involving raters, and covering methodological and applied research. While inter-rater reliability and rater agreement were the top two methods (30.89% [80/259] and 20.08%, respectively), it was found that Rasch measurement theory (including MFRM) and G-theory were the third- and fourth-most employed methods (18.53% and 8.88%, respectively). However, it should be noted that studies using multiple methods were excluded from their analysis. In March 2019, to understand how frequently the two methods have been used together, we electronically searched all past issues of Language Testing (Sage), Language Assessment Quarterly (Taylor and Francis), Assessing Writing (Elsevier), Papers in Language Testing and Assessment (Association for Language Testing and Assessment of Australia and New Zealand: ALTAANZ), and Melbourne Papers in Language Testing (Language Testing Research Centre [LTRC], University of

Melbourne), for relevant studies. We searched for empirical (applied) studies that used the Rasch model and G-theory using keywords "Rasch AND generalizability [generalisability] theory," finding 19 such studies. This number seems small compared to the number of studies in the same five journals from 2010 to 2016 that used only the Rasch model (k = 54; McNamara et al., 2019). As Table 1 shows, the number of studies using both methods has not changed much in language assessment fields over the years. Results suggest that there seems to be much room for the combined use of both methods in language assessment fields.

Some may argue that reporting the results of these methods will take too much space in one journal article; however, results from each method can be reported in separate articles that have different focuses (see Han, 2016, 2019, for examples). Or results from one method could be briefly described in one paper, while those from the other method are extensively reported, depending on the focus of the article (see Fulcher, 1996). Thus, researchers may not need to worry about space limitations when they use the two methods. Another argument against the combined use would be that, depending on the study purpose, the combined use may not be needed. This may certainly be true, but still, we would argue that there are numerous studies that would benefit by using both methods, when conducting basic and advanced test analysis, because each method can produce unique information that the counterpart cannot produce, which helps test developers understand test characteristics and test users confidently interpret and use test scores.

Table 1. Number of studies that used both Rasch model and G-Theory in five language testing journals

Published year	k	Published article
1990–1999	4	Lumley, Lynch, & McNamara (1994), Bachman, Lynch & Mason (1995), Fulcher (1996), Lynch & McNamara (1998)
2000–2009	5	Akiyama (2001), Toyoda & Hashimoto (2001), Kozaki (2004), Sudweeks, Reeve, & Bradshaw (2005), Van Moere (2006)
2000–2018	5	Harsh & Rupp (2011), Hirai & Koizumi (2013), Li & He (2015), Han (2016, 2019)a

Note. k = No. of studies. a = These two articles used G-theory in the analysis, and mentioned that the author used MFRM in the preliminary analysis.

Harsch and Rupp (2011) performed a study using both MFRM and G-theory that is very relevant to our current study. They developed level-specific CEFR-based tasks and analytic rating scales, to assess L2 English writing proficiency from A1 to C1 levels in Germany. Their results were favorable, indicating a close relationship between tasks and the CEFR. They effectively used MFRM and G-theory to highlight different aspects of test takers, tasks, analytic criteria, and raters. Although Harsch

and Rupp (2011) analysed writing tasks, it is useful to compare their work to the current study since the analytical methods were the same and since writing and speaking assessment are similar in terms of eliciting performance using tasks and rating the performance using rating scales.

Like Harsch and Rupp's study (2011), the current study uses both MFRM and G-theory to analyse a CEFR-J spoken interaction test consisting of tasks developed based on CEFR-J descriptors. As with Harsch and Rupp (2011), the tasks are level-specific. However, our levels are more detailed at the lower end of the scale (Pre-A1 to A2.1). We pose three research questions to analyse the tasks at both individual and global levels.

Purposes and Research questions (RQs)

As mentioned above, the purposes of the present study are (a) to examine the measurement quality of CEFR-J spoken interaction tasks aligned with the CEFR-J levels Pre-A1 to A2.1 and (b) to examine the degree to which such tasks align with the CEFR-J levels. Based on these purposes, we pose three RQs.

RQ1: How do test takers, tasks, raters, and a rating scale function in the CEFR-J spoken interaction test?
RQ2: How many tasks are needed to maintain high reliability in seven plausible situations of the CEFR-J spoken interaction test?
RQ3: To what degree is the difficulty of the CEFR-J spoken interaction tasks in line with the difficulty predicted by the CEFR-J levels?

For RQ2, there are seven plausible situations regarding how the CEFR-J tasks are used:

(x) Five one-level situations: Each situation focuses on one CEFR-J level only. The reliability and the number of tasks needed are separately estimated for each level.

(y) One three-level A1 situation: This situation focuses on A1 level only (i.e., A1.1, A1.2, and A1.3). The reliability and the number of tasks needed are estimated for the A1 level.

(z) One five-level situation: This situation focuses on all five levels. The reliability and the number of tasks needed are estimated for all five levels combined.

RQ1 an RQ2 are related to purpose (a), whereas RQ3 is related to purpose (b). RQ1

is answered through MFRM analysis; RQ2 is answered using G-theory; and RQ3 is answered through correlating task difficulty estimates generated by MFRM analysis and CEFR-J difficulty ranks.

Method

Participants

Sixty-six students from a private university and a public university in Japan took a CEFR-J spoken interaction test. Their majors were sports science (n = 49), science and technology (n = 18), and medicine (n = 9). Their L1 was Japanese, and their L2 was English. The median age was 20, and about half of the participants were males (n = 34, 51.52%). Their English proficiency levels ranged from early-beginner to intermediate. They took the test between December 2017 and January 2018, either as part of their classroom activity, or as a paid volunteer.

Although the number of test takers was not large, we considered it sufficient for MFRM analysis based on Linacre (1994a), who states that a sample size of 50 is minimally required for polytomous data, when standard error at 99% level is set at ±1 logits. The sample size of 66 was also considered to be sufficient for G-theory based on Atilgan (2013). Atilgan showed that a sample size of 50 or more leads to small errors in reliability estimates in multivariate G-theory (in a one-facet [p x t] design, to be specific). This method is a more complex one, requiring a larger sample size, than the univariate G-study used in the current research (see e.g., Grabowski & Lin, 2019, for multivariate G-theory).

Spoken interaction test

The CEFR-J spoken interaction test used a face-to-face role-play format, with a teacher (examiner) and a student (test taker) playing each role as specified by a task card (see Introduction for the reasons of selection). Task instructions were given in Japanese. Only English was spoken during the test.

The spoken interaction test was developed based on the CEFR-J descriptors and test specifications we created. It included 10 tasks in total: two tasks per CEFR-J level, with five levels (Pre-A1 to A2.1; 2 tasks x 5 levels = 10 tasks). Each level had two descriptors, each of which was used as the basis for developing a task. Tasks were intended to simulate an authentic situation which secondary school students would likely encounter in their current or future lives. Table 2 shows a brief description of a task situation at each descriptor level. For example, the A2.1.2 task had test takers

engage in an exchange of information while interacting with examiners. This reflects the second descriptor of A2.1, particularly "I can . . . exchange simple opinions, using pictures . . . to help me." Thus, such a situation was selected. The 10 tasks were designed and administered in an increasing difficulty order of the CEFR-J levels.

Table 2. CEFR-J levels, descriptors, and task situations

Level	Task	Task situation	CEFR-J descriptor
Pre-A1	Pre-A1.1	You are in trouble and will ask for help.	I can express my wishes and make requests in areas of immediate need, using basic phrases. I can express what I want by pointing at it, if necessary.
	Pre-A1.2	You will greet your foreign friend on the street.	I can use common, formulaic, daily and seasonal greetings, and respond to those greetings.
A1.1	A1.1.1	You will ask and answer questions about an upcoming event.	I can ask and answer questions about times, dates, and places, using familiar, formulaic expressions.
	A1.1.2	You will talk about your hobbies and ask questions.	I can ask and answer about personal topics (e.g. family, daily routines, hobby), using mostly familiar expressions and some basic sentences (although these are not necessarily accurate).
A1.2	A1.2.1	You will be asked about your favorite sports and hobbies.	I can respond simply in basic, everyday interactions, using a limited repertoire of expressions.
	A1.2.2	You will invite your foreign friend to dinner, asking and answering questions about food that you and your friend like and dislike.	I can exchange simple opinions about familiar topics (e.g. sports, food, likes and dislikes), using a limited repertoire of expressions, provided the other person speaks clearly.
A1.3	A1.3.1	You will be asked about your preference for extracurricular activities, and ask questions about your friend's extracurricular club.	I can make, accept and decline offers, using a limited range of expressions.
	A1.3.2	You will answer questions from your foreign friend who wants to go to the movies together.	I can ask and answer simple questions about very familiar topics (e.g. hobbies, sports, club activities), provided that people speak slowly and clearly with some repetition and rephrasing.
A2.1	A2.1.1	You are a staff member at a tourist office. A tourist will want to know how to get to a few places, and ask you questions. You will give directions using a map.	I can give directions including simple sequencers such as first, then, and next.
	A2.1.2	Using your favorite festival poster, you will describe it (including good and bad points) and ask for opinions.	I can get across basic information and exchange simple opinions, using pictures or objects to help me.

Note. See Tono (2019) for the whole list of descriptors.

Rating scale

A holistic scale of 1 to 3 was used (see Table 3). The three-level scale was selected following national guidelines on classroom assessment of L2 English in Japan (National Institute for Educational Policy Research, 2012); it was also chosen for its simplicity, so that teachers using the scale without much assessment experience can handle it more easily. An analytic scale was not used, as our aim was to relate task achievement with the CEFR-J levels, and a holistic scale allows for direct comparisons.

Table 3. Holistic scale used

Level	Descriptor
3	Achieves the task much better than expected --Shows effortlessness consistently OR --Shows active involvement in the interaction + another strong aspect (e.g., fluency, accuracy)
2	Achieves the task to the degree expected at the given level
1	Fails to achieve the task to the degree expected

To create a holistic scale that aligns with the CEFR-J levels, we first examined CEFR descriptors (Council of Europe, 2001), and watched online videos (e.g., Cambridge Assessment English, 2019; Centre International d'Études Pédagogiques, n.d.) to determine expected performances at A1 and A2 levels. Then we created a scale and discussed divergent scores based on performances in the pilot study, thus clarifying the evaluation criteria.

Test procedures

The test lasted for approximately 30 minutes, including obtaining participants' permission for data use and having them answer a questionnaire. The questionnaire responses were not analysed in this study.

We used two examiners, both of whom were university teachers with substantial experience conducting face-to-face examiner-learner interviews. The test was conducted in a quiet room. Conversation between an examiner and a test taker was voice-recorded in two separate rooms, and videotaped in one of the rooms.

For each task, a card was given to the student. After s/he read the card, 0 to 30 seconds of planning time was allowed, and role play followed for one to three minutes. Planning time and speaking time varied across tasks so as to elicit performances expected from each target level.

Task conditions (such as the duration of speaking time) were decided based on a CEFR-J descriptor and a pilot study. After 10 tasks were created, nine students considered to be at level A1 or higher took the test. On average, each completed three out of 10 tasks. Based on their responses, speaking time was increased, and task instructions were simplified. However, these modifications were very minor, and the tasks were basically the same between the pilot test tasks and the main study tasks. Their scores were used, along with scores from the remaining 57 students.

Scoring

We used two raters, who also served as examiners. The raters, who both held Ph.Ds and were in their forties, were teachers from two different universities and had at least ten-years' experience working as examiners and raters for other speaking tests. In the test administration, an examiner scored performance in each task using a three-level holistic rating scale. Scoring occurred either during or after the completion of a task. Following the test, the raters listened to audio files and double rated approximately 20% of the performances. The two raters discussed the scores and discrepancies until they reached a consensus. However, the current analysis used ratings assigned prior to the discussion.

Analysis

Ratings were analysed using many-facet Rasch measurement (MFRM; Linacre, 1994b and generalizability theory (G-theory; Brennan, 2001). For MFRM, FACETS (Linacre, 2019, Ver. 3.81.2) was used, with three facets included: 66 test takers, 10 tasks, and two raters (736 data points; see Appendix A for the FACETS input). Infit mean squares between 0.50 and 1.50 were considered to fit the Rasch model; less than 0.50 were considered overfit; and more than 1.50 were considered underfit. Because they are typically used since they focus on exemplary patterns in the data, infit mean squares were examined (McNamara et al, 2019). Outfit mean squares were also presented in tables. Appropriateness of the rating scale was judged using the criteria described in Linacre (2002) and Bond and Fox (2015). Bias analysis was also performed to see if there were any biased patterns between test takers and tasks, between test takers and raters, or between tasks and raters.

For G-theory, the data was analysed using the gtheory package in R (Moore, 2016; see Appendix B for the G-theory input) for estimating variance components. We computed percentages, error variance, and reliability using an Excel spreadsheet, not an Excel macro. We employed a fully crossed, balanced, person-by-task (p x t)

design, including tasks as a random facet. In a G-study, a one-facet p x t design partitions total variance into three components: (i) variance between persons (p), (ii) variance between tasks (t), and (iii) a residual variance component (pt, e). The third component combines variance due to an interaction factor (p x t), as well as variance due to unexplained error (e). We estimated reliability according to the number of tasks varied in a D-study. We used a phi coefficient, or index of dependability (Φ), since we intended our tasks to be used for absolute decisions, for instance, to classify test takers into the CEFR-J levels. We did not include a rater facet, and used a score from a single rater (in case of single rating) and an average score from two raters (in case of double rating).

For analysis of the seven situations described in the research question section, we ran separate D-study analyses. We changed a range of included tasks; for example, in a one-level situation (where each CEFR-J level was focused), we treated the two tasks at each level as a single test and employed G-theory. In contrast, in a five-level situation, we treated all 10 tasks as one test.

Results

MFRM results

Global model fit was examined using standardised residuals derived from "unexpected responses" in the output and the criterion established by Linacre (2018): "When the data fit the model, about 5% of standardised residuals are outside ±2, and about 1% are outside ±3" (p. 171). Overall, 3.94% of our standardised residuals went beyond ±2 (29/736), and 2.31% went beyond ±3 (17/736). This suggests that a global model fit was close to adequate. We considered excluding 17 responses that showed large standardised residuals (i.e., 3.00 or more), but decided not to exclude any data and retained all data (data point = 736) for analysis. Although the exclusion of extreme responses would lead to 4.87% (35/719) and 0.70% (5/719) of standardised residuals that went beyond ±2 and ±3 respectively and resulting in an improved model fit, these responses did not show any rating irregularities. Moreover, the exclusion increased underfitting test takers with infit mean squares of 2.00 or more.

The Wright map in Figure 1 shows relationships between test takers, tasks, raters, and a rating scale. Test takers were found to spread more widely than tasks, with raters distributed around zero. As displayed in Table 4, means of test takers, tasks,

and raters were close to each other (0.25, 0.00, and 0.00 logits, respectively). Means of standard errors of test takers, tasks, and raters were within 1 logit. These results broadly suggest that the current test-taker sample fits well with the difficulty of newly developed tasks, and is considered adequate for examining their qualities.

Table 4. Summary statistics of the three facets

	Mean measure	SD measure	Mean SE	SD SE	Separation	Strata	Reliability
Test taker (full marks included)a	0.25	3.01	0.90	0.29	3.01	4.34	.90
Test taker (full marks excluded)b	0.00	2.71	0.87	0.24	2.80	4.07	.89
Task	0.00	2.40	0.32	0.01	7.38	10.17	.98
Rater	0.00	0.23	0.14	0.01	1.23	1.98	.60

Note. SE = Standard error. a = When two test takers who had full marks were included. b = When these two test takers were not included.

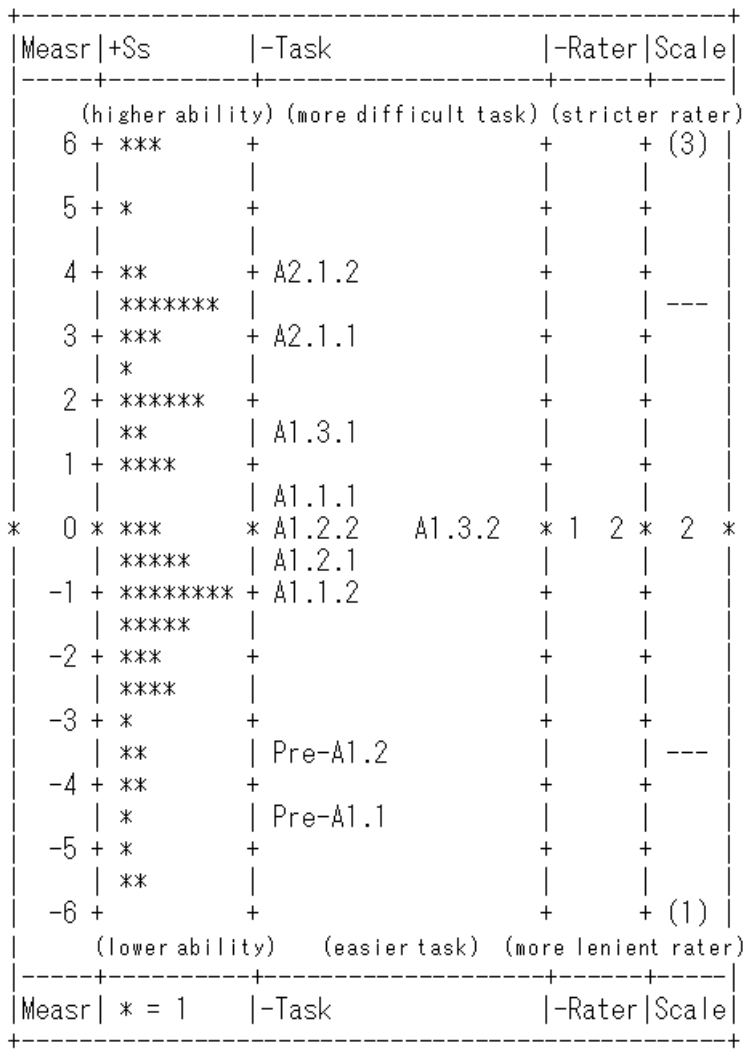

Figure 1. Wright map from the 3-facet analysis.

Test takers

In terms of person fit, 62.12% (i.e., 41/66) of test takers fit the model, having infit mean squares between 0.50 and 1.50; 25.76% of test takers overfit, as shown by their infit mean squares of less than 0.50. There were 11 test takers (16.67%) who had more than 1.50; there were no underfitting test takers with 2.00 or more. Linacre (2018) states that data with infit mean squares of more than 2.0 cause problems in the measurement. The current results showed that there were no such problematic test-taker data. Furthermore, person separation and strata of 3.01 and 4.34 (see Table 4) suggests that test takers were separated into three to four ability levels. Person reliability was high, at .90.

Tasks

All tasks fit the Rasch model, with infit mean squares of 0.74 to 1.31 (see Table 5). This indicates that tasks functioned consistently across test takers and raters. Task separation and strata were 7.38 and 10.17, suggesting that task difficulty could be differentiated into seven to ten levels. Task reliability was high, at .98 (see Table 4).

Table 5. Task measurement report

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Nu	Task
104	71	1.46	1.42	4.00	.35	.86	-.6	.79	-.2	1.12	.51	10	A2.1.2
115	71	1.62	1.68	2.90	.32	1.07	.5	1.06	.2	.91	.43	9	A2.1.1
134	73	1.84	1.92	1.37	.32	1.18	.9	1.22	.8	.84	.47	7	A1.3.1
141	74	1.91	1.96	.73	.31	1.04	.2	.87	-.4	1.02	.33	3	A1.1.1
156	79	1.97	1.99	.22	.31	.76	-1.3	.72	-.9	1.22	.52	6	A1.2.2
159	79	2.01	2.00	-.07	.31	.74	-1.5	.51	-1.8	1.27	.38	8	A1.3.2
152	72	2.11	2.04	-.71	.32	.89	-.6	.79	-.5	1.10	.49	5	A1.2.1
153	73	2.10	2.04	-.80	.33	1.02	.1	1.11	.4	.98	.52	4	A1.1.2
182	75	2.43	2.43	-3.38	.33	1.31	1.4	1.31	.8	.76	.39	2	Pre-A1.2
176	69	2.55	2.64	-4.26	.33	1.17	1.0	1.61	1.1	.75	.34	1	Pre-A1.1
147.2	73.6	2.00	2.01	.00	.32	1.00	.0	1.00	.0		.44		Mean (Count: 10)
23.3	3.1	.31	.32	2.40	.01	.18	1.0	.31	.9		.07		S.D. (Population)
24.6	3.3	.33	.34	2.53	.01	.19	1.0	.33	1.0		.08		S.D. (Sample)

Raters

Both raters fit the Rasch model, with infit mean squares of 0.99 and 1.01 (see Table 6). These results suggest that they provided consistent ratings across test takers and tasks. There was also high interrater agreement (75.30%), which was greater than predicted by Rasch analysis (74.60%). The two raters' severity estimates were very close (0.23 and −0.23). Rater separation and strata were 1.23 and 1.98. Rater reliability was not high, at .60 (see Table 4). These results suggest that raters behaved similarly, and this is probably because they developed the tasks and the scale collaboratively, went through rater training, and had similar assessment and teaching backgrounds. Additionally, performance evaluation may have been easier in the current test than in other speaking tests, as the tasks were easy and the test-takers' output was not very extensive.

Table 6. Rater measurement report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	Exact Agree. Obs %	Exact Agree. Exp %	N	Rater
817	411	1.99	1.99	.23	.14	.99	-.1	1.00	.0	1.03	.48	75.3	74.6	2	2
655	325	2.02	2.01	-.23	.15	1.01	.1	.98	.0	.99	.52	75.3	74.6	1	1
736.0	368.0	2.00	2.00	.00	.14	1.00	.0	.99	.0		.50			Mean (Count: 2)	
81.0	43.0	.01	.01	.23	.01	.01	.2	.01	.1		.02			S.D. (Population)	
114.6	60.8	.02	.02	.33	.01	.02	.2	.02	.1		.03			S.D. (Sample)	

Rating scale analysis

In terms of rating scale functioning, four out of five criteria were met. First, each level (from 1 to 3) had more than 10 data counts (121 to 462 in the "used category counts"; see Table 7). Second, the average measures for test takers at a given level increased as levels increased (−4.59, −0.07, and 4.51; see "average measures" in the "quality control"). Rasch-Andrich threshold measures (called step difficulty, or step calibration) also increased as levels increased (−3.67 and 3.67). Third, the probability curve of the scale had a clear top (see Figure 2). Fourth, the fit statistics (outfit mean squares) were less than 2.0 (see "Outfit MnSq" in the "quality control"). Fifth, the distance between step calibrations (or the difference between Rasch-Andrich threshold measures) was 7.34 (3.67 − (−3.67)), which was larger than 5.0 logits. This result indicates that level boundaries are too distant to obtain sufficient information and precision (Linacre, 2002). In other words, Level 2 was too easy to achieve and/or Level 3 was too difficult to achieve, and Level 2 did not provide sufficient information. Based on this result, we discussed how to modify the scale (including scale descriptors and rater training) to make Level 2 more difficult and/or Level 3 easier. We did not consider increasing the number of levels because of the Japanese national guidelines on L2 English classroom assessment (National Institute for Educational Policy Research, 2012) and because of the usability of the scale for less-experienced teachers.

Table 7. Category statistics of the rating scale

DATA Score	Category Counts Total	Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	RASCH-ANDRICH Thresholds Measure	S.E.	EXPECTATION Measure at Category -0.5	from	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
1	137	137	19%	19%	-4.59	-4.56	1.0			(-4.73)		low	low	100%
2	462	462	64%	83%	-.07	-.09	1.1	-3.67	.14	.00	-3.64	-3.67	-3.66	95%
3	137	121	17%	100%	4.51	4.52	1.0	3.67	.15	(4.74)	3.65	3.67	3.66	100%
										(Mean)		(Modal)	(Median)	

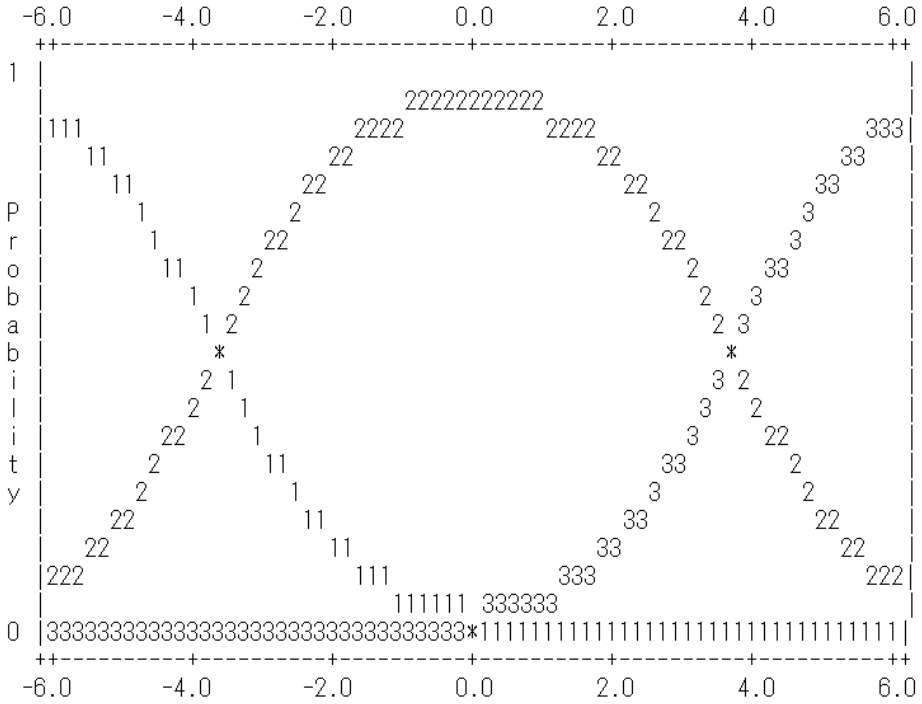

Figure 2. Probability curves of the rating scale.

Bias analysis

Table 8 shows the results for bias analysis, which was conducted separately between test takers and tasks, between test takers and raters, and between tasks and raters. The percentages of large t values (standardised bias scores) were low across all analyses. This suggests only a limited percentage of bias interactions. Some biased responses were examined and used for further improvement in the speaking test.

Table 8. Summary statistics for bias analysis

	Test taker x Task	Test taker x Rater	Task x Rater
No. of interactions	578	83	20
Absolute t values ≥ 2	18 (3.11%)	4 (4.82%)	0 (0.00%)
Maximum t	3.18	2.65	1.41
Minimum t	–2.93	–2.66	–1.30
M	0.03	–0.01	–0.01
SD	0.83	0.74	0.84

Note. No. of interactions = Number of combinations that exist between test takers by tasks, test takers by raters, and tasks by raters.

Relationship between difficulty of the test tasks and CEFR-J levels

To answer RQ3, we correlated empirical task difficulty estimates from the MFRM and CEFR-J difficulty ranks (e.g., with Pre-A1.1 and Pre-A1.2 coded as the same) by using the Spearman rank-order correlation coefficient (ϱ). There was a strong relationship (ϱ = .81) between the two. Overall, this suggests a piece of validity evidence that task difficulty in the current test is in line with the CEFR-J levels. However, the correlation was not perfect, because, as Figure 1 showed, (a) the A1.1.1 task was too difficult (with a difficulty level similar to the A1.3.1 task, and (b) the A1.3.2 task was too easy (with a difficulty level similar to the A1.2.2 task). These results are line with Harsch and Rupp (2011), who found a strong agreement between actual task difficulty estimates and difficulty levels predicted from the CEFR levels. They also reported that one task constructed to be at the A1 level turned out to have a similar difficulty as the A2 level. The CEFR-J classifies the A1 level into three sublevels. We can expect that differentiating these three levels would be more difficult than differentiating the A1 and A2 levels. We discussed this issue to revise the problematic tasks, so that our CEFR-J tasks align better with the CEFR-J descriptors in terms of difficulty.

G-theory results

Table 9 shows results in seven situations. For example, in a one-level Pre-A1 situation, (iii) a residual variance accounted for the largest percentage (62.78%[2]), followed by (i) a person variance (35.78%) and then (ii) a task variance (1.45%), in a situation where a single task was used.

[2] The percentage of a residual variance (62.78%) was calculated using unrounded numbers (i.e., 0.1691936366153450/[0.0964246877850972 + 0.0038952454553425 + 0.1691936366153450]), while the use of rounded numbers would lead to a slightly different percentage (62.83%; 0.169/[0.096 + 0.004 + 0.169] = .6283).

Table 9. G-Study variance components (and percentages) of the test and reliability estimates in seven situations

Level	One					Three	Five
	Pre-A1	A1.1	A1.2	A1.3	A2.1	A1.1 to A1.3	Pre-A1 to A2.1
Persons (p)	0.096 (35.78)	0.125 (38.49)	0.195 (66.97)	0.090 (36.79)	0.208 (63.62)	0.130 (45.76)	0.132 (32.31)
Tasks (t)	0.004 (1.45)	0.024 (7.26)	0.007 (2.30)	0.009 (3.78)	0.008 (2.51)	0.008 (2.94)	0.117 (28.55)
Residual (pt, e)	0.169 (62.78)	0.177 (54.24)	0.090 (30.74)	0.145 (59.43)	0.111 (33.87)	0.146 (51.30)	0.160 (39.14)
k	2	2	2	2	2	6	10
Error	0.087	0.100	0.048	0.077	0.059	0.026	0.028
Φ	.53	.56	.80	.54	.78	.84	.83

Note. k = Number of tasks. Error = Absolute error variance, which was calculated using "Task variance/k + Residual variance/k" (e.g., 0.004/2 + 0.169/2 = 0.002 + 0.085 = 0.087 in a one-level Pre-A1 situation). Φ = Phi coefficient, calculated using "Person variance/(Person variance + Absolute error variance)" (e.g., 0.096/(0.096 + 0.087) = .53 in a one-level Pre-A1 situation).

In seven situations, we found that the high variance contribution was due to (a) persons in a one-level A1.2 situation (66.97%), (b) persons in a one-level A2.1 situation (63.62%), and (c) residuals in the other situations (e.g., a five-level situation: 39.14%). Interestingly, task variance tended to be small in one-level situations (1.45% to 7.26%) and in the three-level A1 level situation (2.94%). However, task variance was quite large in the five-level situation (28.55%). This suggests that tasks at one CEFR-J level, and even at the A1 level (with A1.1 to A1.3 combined), are similar in difficulty. This means that Pre-A1, A1, and A2.1 tasks are of three distinct difficulty levels (as defined in the CEFR-J and operationalised in the current study). They are arranged in order of increasing difficulty. Notably, when Pre-A1, A1, and A2.1 tasks are grouped into a test, task variance increased. Since different difficulty levels were expected, the similarity in difficulty from A1.1 to A1.3 tasks may be problematic. Further revision is necessary to differentiate these three levels.

The last row in Table 9 shows reliability estimates in different situations. When Φ = .70 was used as a criterion, high reliability was observed in the one-level A1.2 situation (.80), one-level A2.1 situation (.78), three-level A1.1 to A1.3 situation (.84), and five-level Pre-A1 to A2.1 situation (.83). The reliability in the one-level Pre-A1, A1.1, and A1.3 situations were low (.53, .56, and .54, respectively). In general, reliability tended to be higher as the number of tasks included increased. However, even with two tasks, reliability in two situations (i.e., the one-level A1.2 situation and the one-level A2.1 situation) were high (.80 and .78, respectively). This is because

reliability generally increases when task variability is smaller and person variability is larger. In other words, when tasks can clearly differentiate test-takers' ability levels, higher reliability can be derived. This should be kept in mind in developing tasks; we should aim to create tasks that can well separate learners according to different ability levels.

When reliability is not very high, it can be increased by raising the number of tasks in the test. Figure 3 shows how reliability changes according to the number of tasks. The current CEFR-J spoken interaction tasks could be used in both low-stakes and high-stakes situations. Accordingly, three criteria for high reliability were set (Φ = .70, .80, and .90) to accommodate a range of decisions, from low-stakes to higher-stakes (Wells & Wollack, 2013). Table 10 shows the specific number of tasks necessary to obtain reliability of .70, .80, and .90. It was found that, for example, in the one-level A1.2 situation, the number of tasks required was two, two, and five, respectively. The number of tasks required was small in one-level A1.2 and one-level A2.1 situations, where person variance was large and task variance was small. The opposite was true in situations where person variance was small and task variance was large (for example, in the five-level situation, where five, nine, and 19 tasks were necessary to obtain reliability of .70, .80, and .90, respectively).

Table 10. Number of tasks needed to obtain high reliability estimates in seven situations

Level	One					Three	Five
	Pre-A1	A1.1	A1.2	A1.3	A2.1	A1.1 to A1.3	Pre-A1 to A2.1
Φ = .70	5	4	2	4	2	3	5
Φ = .80	8	7	2	7	3	5	9
Φ = .90	17	15	5	16	6	11	19

Note. "Φ = .70" = No. of tasks that could be required to get Φ = .70

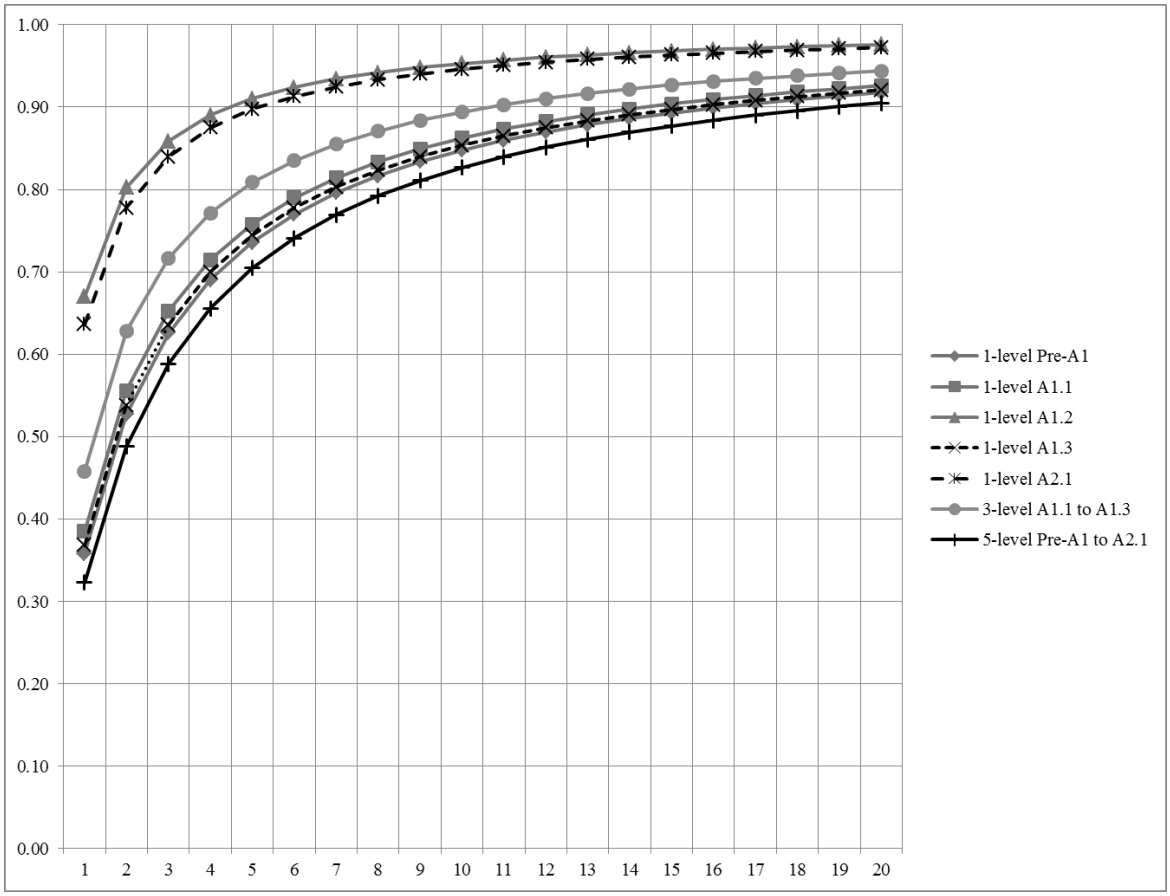

Figure 3. Changes in reliability according to the number of tasks included. X axis = Number of tasks included. Y axis = Φ.

Discussion and conclusions

This study examined the measurement quality of the CEFR-J spoken interaction tasks designed for Pre-A1 to A2.1 levels. RQ1 asked how test takers, tasks, raters, and a rating scale function in the CEFR-J spoken interaction test. Overall the MFRM results were generally positive. However, there were two points that did not meet the criteria. First, there was a relatively high percentage of underfitting test takers. Second, the distance between step calibrations in the rating scale was large.

RQ2 asked how many tasks are needed to maintain high reliability in seven plausible situations of the CEFR-J spoken interaction test, using G-theory. Using Φ = .70 as a criterion, the number of tasks required ranged from two to five. Depending on the situation in which the test is used, and how precise ability assessment is required to be, the expected number of tasks varied.

RQ3 examined to what degree is the difficulty of the CEFR-J spoken interaction tasks in line with the difficulty predicted by the CEFR-J levels. The high correlation between expected and actual difficulty (ϱ = .81) suggests a strong degree of agreement between the two. Nevertheless, there were two tasks (i.e., A1.1.1 and A1.3.2) that did not conform to the expected levels. These results suggest the need to examine and revise the tasks and the rating scale, as well as to improve rater training.

Both MFRM and G-theory provided distinctive information from different perspectives. MFRM showed detailed, specific measurement characteristics in the data such as task and rater fit and rating scale functioning. In contrast, G-theory presented simulation results in terms of how the number of tasks changes reliability in plausible situations.

However, there were also results from similar perspectives. First, test-taker reliability in MFRM (.90) and reliability in the five-level situation in G-study (.83) were similar. Second, results of a task facet in G-study suggest that task contributions in one-level situations are smaller than those in the five-level situation (up to 7.26% vs. 28.55%). Thus, task variation was much larger between tasks at different CEFR-J levels than between tasks within a particular level. This result aligns well with findings from the Wright map in MFRM: tasks across the five CEFR-J levels have a wide logit range overall (i.e., 8.26 = 4.00 − (−4.26)). However, tasks within the same level tend to cluster together visually, as well as in difficulty estimates, with the largest range being 1.53 at A1.1 level (i.e., 0.73 − (−0.80)). Third, as described in the G-study results section, A1-level tasks had small task variances, and thus similar difficulties. The Wright map in MFRM also shows six tasks at A1 level (A1.1. to A1.3 levels) forming one cluster. This seems to suggest that there are three task groups, each of varying difficulty. However, task separation and strata from MFRM were 7.38 and 10.17, indicating larger task groups. This underscores the importance of examining data from various angles.

In addition to the areas for improvement suggested above, three additional points should be improved in future examination of the revised tasks. First, we had only two tasks at each CEFR-J level. This needs to be expanded, increasing the task bank's volume, given the G-theory results that more than two tasks are usually needed to provide sufficiently precise assessment. Doing so would enable wider coverage of spoken interaction ability, as well as more precise estimates in MFRM and G-theory. Second, all participants in this study were university students. It is necessary to recruit a wider population of students (including secondary school pupils in Japan)

to examine how a representative body of test takers responds to the tasks. Third, along with the revision of tasks from Pre-A1 to A2.1 levels, tasks at higher levels and a manual for developing CEFR-J tasks will be developed and trialed. All resources will be made publicly available. Fourth, G-theory should include a rater facet, which helps researchers examine interactions between test takers, tasks, and raters separately. Further, while we used a role-play format in which the test taker and examiner interact, there are alternative formats (e.g., peer-to-peer interactive formats). Comparing different formats for assessing spoken interactions using MFRM and G-theory would shed new light on this type of assessment.

Overall, the current study shows the usefulness of MFRM and G-theory, especially when used in combination. The detailed views they provide enable researchers and test users to discover a test's strengths, weaknesses, and other characteristics; due to the complex interplay between test takers, tasks, raters, and rating scales, the combined use of MFRM and G-theory is particularly useful in assessing spoken interaction. Such examples are problems with the rating scale's levels and tasks that appeared in an unexpected order of difficulty (identified in MFRM) and low reliability in plausible one-level situations (identified in a G-study in G-theory). Moreover, a D-study in G-theory can estimate the minimum number of tasks required. This is hypothetical in terms of considering tasks not yet extant, but would surely be useful in striking a balance between the desire to include more tasks, and the necessity of limiting their numbers. Many other useful methods exist, and their selection should be guided by investigators' research focuses. Clearly, though, an effective selection of multiple methods (including MFRM and G-theory) would lead to better illumination of test characteristics. Further, the current study suggests the importance of conducting rigorous preliminary analysis to revise tasks and rating procedures before a test is operationalised. Although this sounds intuitive, it is not always followed in practice. In cases where pilot studies are not possible, including more tasks would allow researchers to select from a wider range of appropriate tasks for ability estimation and test use. This is essential, particularly for tasks that are based on frameworks such as the CEFR-J when task difficulty is expected to align with established proficiency levels.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (A), No. 16H01935. We would like to thank Masashi Negishi, Yukio Tono, and other research team members for their

guidance. This article is based on two presentations. The first was delivered at the Pacific-Rim Objective Measurement Symposium (PROMS) 2018, Fudan University, Shanghai, China on July 25, 2018. The second was presented at the CEFR-J 2019 Symposium, Doshisha University, Kyoto, Japan on March 23, 2019.

References

Akiyama, T. (2001). The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom context. *Melbourne Papers in Language Testing*, *10*, 1–21.

Atilgan, H. (2013). Sample size for estimation of G and Phi coefficients in generalizability theory. *Eurasian Journal of Educational Research*, *51*, 215–227. Retrieved from https://eric.ed.gov/?id=EJ1059904

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*, 238–257. doi:10.1177/026553229501200206

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. III: Evaluation, Methodology, and Interdisciplinary Themes, Part 10: Quantitative analysis, pp. 1301–1322). West Sussex, UK: John Wiley & Sons.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Brennan, R. L. (2001). Generalizability theory. New York, NY: Springer.

Cambridge Assessment English. (2019). *Fitness for purpose: Examples of speaking tests.* Retrieved from http://www.cambridgeenglish.org/research-and-validation/fitness-for-purpose/

Centre International d'Études Pédagogiques. (n.d.). DVD: *Spoken performances illustrating the 6 levels of the Common European Framework of Reference for Languages.* Retrieved from http://www.ciep.fr/en/books-and-cd-roms-dealing-with-assessment-and-certifications/dvd-spoken-performances-illustrating-the-6-levels-of-the-common-european-framework-of-reference-1

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.

Council of Europe. (2004). *Gaikokugo kyoiku II: Gaikokugo no gakushu, kyoju, hyoka notameno yoroppa kyotsu sanshowaku* [Foreign language education II: Common European Framework of Reference for Languages for foreign learning,

teaching, and assessment] (Yoshijima, S., & Ohashi, R., Trans.). Tokyo: Asahi Press. (Original work published 2001)

Deygers, B., Van Gorp, K., & Demeester, T. (2018). The B2 level and the dream of a common standard. *Language Assessment Quarterly*, *15*, 44–58. doi:10.1080/15434303.2017.1421955

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd revised and updated ed.). Frankfurt am Main, Germany: Peter Lang.

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment* (Vol. I: Fundamental techniques; pp. 153–176). New York, NY: Routledge.

Engelhard, Jr. G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales.* New York, NY: Routledge.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*, 208–238. doi:10.1177/026553229601300205

Galaczi, E., & ffrench, A. (2011). Context validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 112–170). Cambridge, UK: Cambridge University Press.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15,* 219–236. doi:10.1080/15434303.2018.1453816

Gebril, A. (2013). Generalizability theory in language testing. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. West Sussex, U.K.: John Wiley and Sons. doi:10.1002/9781405198431.wbeal1326

Grabowski, K. C., & Lin, R. (2019). Multivariate generalizability theory in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment* (Vol. I: Fundamental techniques; pp. 54–80). New York, NY: Routledge.

Han, C. (2016). Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach. *Language Assessment Quarterly, 13*, 186–201. doi:10.1080/15434303.2016.1211132

Han, C. (2019). A generalizability theory study of optimal measurement design for a summative assessment of English/Chinese consecutive interpreting.

Language Testing, 36, 419–438. doi:10.1177/0265532218809396

Harsch, C. (2018). How suitable is the CEFR for setting university entrance standards? *Language Assessment Quarterly*, *15*, 102–108. doi:10.1080/15434303.2017.1420793

Harsch, C., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, *8*, 1–33. doi: 10.1080/15434303.2010.53557

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a Story Retelling Speaking Test. *Language Assessment Quarterly*, *10*, 398–422. doi:10.1080/15434303.2013.824973

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing*, *16*, 163–188. doi:10.1177/026553229901600203

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, *21*, 1–27. doi:10.1191/0265532204lt272oa

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly*, *12*, 178–212. doi:10.1080/15434303.2015.1011738

Linacre, J. M. (1994a). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions, 7*, 328. Retrieved from https://www.rasch.org/rmt/rmt74m.htm

Linacre, J. M. (1994b). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

Linacre, J. M. (2018). *A user's guide to FACETS Rasch-model computer programs: Program manual 3.81.0.* Retrieved from http://www.winsteps.com/manuals.htm

Linacre, J. M. (2019). FACETS: *Many-Facet Rasch-measurement* (Version 3.81.2) [Computer software]. Chicago: MESA Press.

Lumley, T., Lynch, B. & McNamara, T. (1994). A new approach to standard-setting in language assessment. *Papers in Language Testing and Assessment*, *3*(2), 19–40.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*, 158–180. doi:10.1177/026553229801500202

McNamara, T., Knoch, T., & Fan, J. (2019). *Fairness, justice, and language assessment.* Oxford, UK: Oxford University Press.

Moore, C. T. (2016). *gtheory: Apply generalizability theory with R*. Retrieved from https://cran.r-project.org/web/packages/gtheory/index.html

National Institute for Educational Policy Research. (2012). *Hyoka kijun no sakusei, hyoka hoho to no kuhu kaizen no tameno sanko shiryo (koko gaikokugo)* [Reference documents for Japanese senior high school foreign language studies for the development of assessment criteria and improvement of assessment methods and others]. https://www.nier.go.jp/kaihatsu/hyouka/kou/11_kou_gaikokugo.pdf

Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In E. D. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks: Proceedings of the ALTE Kraków Conference* (pp. 135–163). Cambridge, UK: Cambridge University Press.

O'Sullivan, B., & Green, A. (2011). Test taker characteristics. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 36–64). Cambridge, UK: Cambridge University Press.

Papageorgiou, S. (2016). Aligning language assessments to standards and frameworks. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 327–340). Berlin, Germany: De Gruyter.

Sawaki, Y., & Xi, X. (2019). Univariate generalizability theory in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment* (Vol. I: Fundamental techniques; pp. 30–53). New York, NY: Routledge.

Schmidt, M. G., Runnels, J., & Nagai, N. (2017). The past, present and future of the CEFR in Japan. In F. O'Dwyer, M. Hunke, A. Imig, N. Nagai, N. Naganuma, & M. G. Schmidt (Eds.), *Critical, constructive assessment of CEFR-informed language teaching in Japan and beyond* (pp. 18–48). Cambridge, UK: Cambridge University Press.

Shiina, K. (2013). CEFR wa hyoukanotame? Shido notame? [Should CEFR be used for assessment or teaching?] In Y. Tono (Ed.), *Can-Do risuto sakusei katsuyo eigo totatudo shihyo CEFR-J gaidobuku* [The CEFR-J handbook: A resource book for using Can-Do descriptors for English language teaching] (pp. 51–55). Tokyo: Taishukan.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of

college sophomore writing. *Assessing Writing*, *9*, 239–261. doi:10.1016/j.asw.2004.11.001

Tono, Y. (2019). *CEFR-J.* Retrieved from http://www.cefr-j.org/

Toyoda, E., & Hashimoto, Y. (2001). Analysis of a new Japanese language placement test battery using G-theory and Rasch model programs. *Melbourne Papers in Language Testing, 10*, 49–66.

University of Cambridge ESOL Examinations. (2011). *Using the CEFR: Principles of good practice.* Cambridge, UK: Cambridge ESOL.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, 23, 411–440. doi:10.1191/0265532206lt336oa

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of statistics, 26*, 81–124.

Wells, C. S., & Wollack, J. A. (2013). *An instructor's guide to understanding test reliability.* Madison, WI: Testing & Evaluation Services, University of Wisconsin. Retrieved from https://testing.wisc.edu/instructionalsupport.html

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*, 161–192. doi:10.1177/0265532216686999

Appendix A. FACETS input

```
title = 190401CEFRJ_ST_N66
convergence = 0.1 ; size of largest remaining marginal score residual at convergence
unexpected = 2 ; size of smallest standardized residual to report
arrange = m ; arrange output tables in Num decending and Logit ascending order
facets = 3 ; 3 facets    1 Person, 2 Task, 3 Rater
noncenter = 1 ; examinee facet floats
positive = 1 ; for examinees, greater score      greater measure
Pt-biserial = Yes ; report the point-biserial correlation
Inter-rater = 3 ; facet 3 is the rater facet
Yardstick = 0,2,-6,6
Missing= N
Model=
?B,?B,?,R3
?B,?,?B,R3
?,?B,?B,R3
*

Labels=
1,Ss
1 = 001
(omitted)
66 = 066
*

2,Task
1 = Pre-A1.1
2 = Pre-A1.2
3 = A1.1.1
4 = A1.1.2
5 = A1.2.1
6 = A1.2.2
7 = A1.3.1
8 = A1.3.2
9 = A2.1.1
10 = A2.1.2
*
```

```
3,Rater
1-2
*

data =
001 1-10 1 3 N 2 N 3 N N N N N
  (omitted)
066 1-10 2 2 2 1 1 2 2 2 2 1 1
```

Appendix B. G-theory input

```
# TITLE: G-Study of CEFR Speaking Data (p x t)

datafolder <- "C:/Users/-/Desktop/CEFR Gstudy Jan2019"
setwd(datafolder)

# Run G studies
library(gtheory)
formula1 <- "Score ~ (1 | Person) + (1 | Task)"

# Five-level PreA1-A2.0 design
Model1 <- gstudy(datalong_subset5, formula = formula1)

# Three-level A1all design
datalong_subsetA11A12A13 <- datalong_subset5[which(datalong_subset5$Task != 1 & datalong_subset5$Task != 2 & datalong_subset5$Task != 9 & datalong_subset5$Task != 10),]

Model2 <- gstudy(datalong_subsetA11A12A13, formula = formula1)

# One-level PreA1 & A1.1 & A1.2 & A1.3 & A2.1 design
datalong_subsetPreA1 <- datalong_subset5[which(datalong_subset5$Task == 1 | datalong_subset5$Task == 2),]
datalong_subsetA11 <- datalong_subset5[which(datalong_subset5$Task == 3 | datalong_subset5$Task == 4),]
datalong_subsetA12 <- datalong_subset5[which(datalong_subset5$Task == 5 | datalong_subset5$Task == 6),]
datalong_subsetA13 <- datalong_subset5[which(datalong_subset5$Task == 7 | datalong_subset5$Task == 8),]
datalong_subsetA21 <- datalong_subset5[which(datalong_subset5$Task == 9 | datalong_subset5$Task == 10),]

Model3a <- gstudy(datalong_subsetPreA1, formula = formula1)
Model3b <- gstudy(datalong_subsetA11, formula = formula1)
Model3c <- gstudy(datalong_subsetA12, formula = formula1)
Model3d <- gstudy(datalong_subsetA13, formula = formula1)
Model3e <- gstudy(datalong_subsetA21, formula = formula1)
```

```
# Generate csv output tables for variance components and percents
varcomponents <- rbind(t(Model1$components$var),
                       t(Model2a$components$var),
                       t(Model2b$components$var),
                       t(Model3a$components$var),
                       t(Model3b$components$var),
                       t(Model3c$components$var),
                       t(Model4a$components$var),
                       t(Model4b$components$var),
                       t(Model4c$components$var),
                       t(Model4d$components$var),
                       t(Model5a$components$var),
                       t(Model5b$components$var),
                       t(Model5c$components$var),
                       t(Model5d$components$var),
                       t(Model5e$components$var))

write.csv(varcomponents, file = "varcomponents.csv", row.names = FALSE)
```