

Multifaceted Rasch Analysis of Paired Oral Tasks for Japanese Learners of English

Rie Koizumi, Yo In'nami and Makoto Fukazawa

Introduction

The use of multifaceted Rasch measurement (MFRM) has prevailed in the assessment field, especially in assessing second language (L2) speaking and writing, which involves complex interactions between test takers, tasks, raters, rating scales, and other factors. McNamara and Knoch (2012) describe how Rasch measurement, especially MFRM was adopted in L2 testing communities in the 1990s. Recent applications include Davis (2016) and Aryadousta (2016), both of which investigated the complex nature of L2 speaking assessment.

One type of L2 speaking assessment that has attracted attention from teachers and test developers is a paired oral test (paired oral, hereafter). In paired orals, test takers make pairs and talk with each other rather than an interviewer, and interactions are evaluated by raters. The English ability to interact with various speakers, convey facts precisely, and express one's opinions promptly, while responding to listeners and managing interaction should be fostered and measured, since this ability is one of the determinants of success in today's globalized world. Although there are benefits to using paired orals in L2 speaking assessment, paired oral-related research and applications are limited in Japan. Koizumi et al. (in press) developed and examined a paired oral test consisting of four tasks for Japanese university students learning L2 English. This study builds on Koizumi et al. and expands the number of paired oral tasks calibrated on a logit scale and examines its usefulness.

R. Koizumi (✉)
Juntendo University, Chiba, Japan
e-mail: rkoizumi@juntendo.ac.jp

Y. In'nami
Chuo University, Tokyo, Japan

M. Fukazawa
University of the Ryukyus, Okinawa, Japan

Previous Studies on Paired Orals

The literature has shown that paired orals have unique values different from other types of speaking assessment with teacher–candidate interaction, especially in two points: First, paired orals can measure oral interaction that is likely observed in natural, real-life conversation when conversation partners have equal status, because the test takers have chances and are required to be active in maintaining the conversation and producing the discourse in cooperation with another test taker of an equal status (e.g., Galaczi and French 2011). Second, paired oral formats are usually similar to pair activities often conducted in communicative classes. Using paired orals can give students a clear message that what they are doing in class is important for good grades and eventually for their future use of English. Similarities between teaching and assessment activities also make it easier for teachers to relate the assessment results for instruction (e.g., Negishi 2015).

Paired orals have two main disadvantages (Negishi 2015). Firstly, test takers' performance and scores can be affected by factors other than their L2 proficiency, such as their own and their partners' L2 proficiency, personality, and degree of familiarity with each other (e.g., Galaczi and French 2011). Second, paired orals tend to have lower reliability across raters and test occasions than examiner–interview and monolog formats, as can be inferred from the study of a group oral test with four test takers discussing topics (Van Moere 2006). Although these are certainly issues that need to be addressed, they may not matter much in low-stakes testing contexts such as classroom assessment, where teachers can assess and judge students' ability on multiple occasions in combination with a few formats.

Paired orals have been examined from the perspective of factors affecting test scores (e.g., Davis 2009; Galaczi 2008, 2014; Kley 2015; Taylor and Wigglesworth 2009, special issue in *Language Testing*) and incorporated into major speaking tests such as the Cambridge English exams (Galaczi and French 2011). However, in Japan, their research and applications are limited, with a few exceptions such as Negishi (2015) and Koizumi et al. (in press). Negishi (2015) compared university students' performances across three formats (i.e., picture description, paired oral, and group oral) using MFRM. She reported that all test formats and raters fit the Rasch model, the paired oral was the most difficult, followed by the picture description, and the group oral, in that order, and test formats and test takers' proficiency levels affected scores. Koizumi et al. (in press) developed a paired oral test and investigated the validity of the interpretation of paired oral test scores, from four viewpoints: First, all tasks and raters fit the Rasch model, with appropriate rating scale properties. Second, high reliability was observed with one task and two raters, or with three tasks and a single rater (with the cutoff score of $\varphi = 0.70$). Third, the test had a unidimensional structure of one factor affecting all scores. Fourth, paired oral scores were moderately correlated with scores of the Test of English as a Foreign Language (TOEFL) Institutional Testing Program (ITP), as predicted in the test development stage.

Current Study

Considering benefits of paired orals, especially in L2 classroom assessment, but limited applications in Japan, it is important to make them accessible to teachers in Japan. For this purpose, we create a bank of tasks whose difficulty levels are measured with a practical rating scale and whose features related to task are set based on Koizumi et al. (in press). To the authors' knowledge, there are no publications describing an attempt to create a large task bank for paired orals in and outside Japan.

The current study aims to assess the L2 oral interactive ability of university students at the novice and intermediate levels. Using MFRM as well as structural equation modeling (SEM) and generalizability theory, we examine the following six questions that are associated with aspects of validity (Messick 1996) and inferences required to make a plausible validity argument (Chapelle et al. 2008). These six research questions (RQs) and validity aspects and inferences are shown below in the parentheses. The current study examines essential aspects of validity in order to provide building blocks of evidence for validity of the interpretation and use based on paired oral test scores.

1. Does the test have a unitary factor structure underlying the paired oral? (structural aspect; Explanation inference)
2. Do all tasks and raters fit the Rasch model? (content and structural aspects; Evaluation and Generalization inferences)
3. Do test tasks have a wide range of difficulty and no wide gaps in difficulty? (content and structural aspects; Evaluation inference)
4. Is the difficulty of student cards equal? (generalizability aspect; Generalization inference).
5. Does the holistic rating scale function properly? (structural aspect; Evaluation inference)
6. How many tasks and raters are minimally needed to obtain sufficient reliability? (generalizability aspect; Generalization inference).

RQ1 examines the test structure of the paired oral; assessing only a single dimension (unidimensionality) is a crucial assumption for Rasch analysis to be met. RQ2 and RQ5 address qualities of tasks, raters, and the holistic rating scale used in this study. RQ3 examines the distribution of tasks to determine whether the tasks are sufficient in number and range for assessing novice- and intermediate-level learners of English. We argue that a task bank should have many tasks with a wide range of difficulty and should not have no-task areas on the Rasch logit scale, since we intend to create a bank of tasks useful for teachers to choose from depending on their test purposes and on their target learners who may have different levels of ability. RQ4 examines comparability of student cards. In the paired oral, one student receives a student card either for Student A or B. We intend to make each card's difficulty level equal and examine this in the RQ. RQ6 inspects the degree of reliability that the paired oral can assure depending on the number of tasks and raters.

Method

Participants

A total of 190 students from three private universities in Japan participated. Their majors were technology, medicine, or English. Their L2 proficiency levels ranged mostly from novice to intermediate. Most of the participants were originally from Japan, so their mother tongue was Japanese. A majority were first-year students who had studied English for at least 6 years at secondary school. Some students were from other countries, but we included them because we intended to create a test for classrooms at Japanese universities that have some overseas students. In their English lessons, the students were instructed to make a pair by themselves to mitigate the influence of familiarity. We did not control or examine the effect of proficiency this time but this should be addressed in the future.

Materials and Procedures

The test included an easy warm-up task and 11 assessment tasks—seven role-plays and four discussions. We used four tasks (Tasks 1–4) similar to the ones used in Koizumi et al. (in press) but modified some instructions by providing specific contexts for the conversation and more familiar place names, and created seven new tasks with familiar topics (see Table 1 for all the tasks). The students were requested to talk for about 2 or 3 min per task. They were not given any planning time and were encouraged to talk in a natural, two-way style with back channeling and eye contact.

After making a pair, students received a student card either for Student A or B, which provided a warm-up and 11 tasks. For raters to identify who was speaking, students were told to begin each task with their name. Tasks were either role-play or discussions. In the role-play task, the card contained a role to play and who should speak first (see Table 1). For example, in Task 9 (Role-play 5: Toothache) Student A needs to begin the conversation, and say that s/he has a terrible toothache; Student B should respond with sympathy and suggest going to see a dentist or take a painkiller; Student A should refuse suggestions at least once and they should continue the conversation. Out of seven role-play tasks, Student A should begin in three, whereas Student B should do so in the remaining four. We intended to make the Student A and B cards comparable in terms of difficulty.

The order of performing tasks was partially counterbalanced: Approximately a third of students performed a warm-up task and tasks 1–10 (not 11) in that order. Another third performed a warm-up task, tasks 4–10, and 1–3 (not 11) and the rest performed a warm-up task, tasks 8–10, and 1–7 (not 11). One class performed only tasks 3, 5, 8, 6, 9, 7, 10, and 11, in that order, because of the limited class time.

Table 1 Warm-up tasks and 11 assessment tasks used

Task	Instruction
Warm-up [A]	Talk about (a) brothers or sisters, (b) pets, (c) boys' (or girls') high school or coeducational school, or (d) favorite food (2 min)
1. Club (RP1) ^a [B]	For A: <u>You are in a cooking club. B is considering whether to join it and wants to ask you questions. Use the information below and answer B's questions kindly. <i>New Cooking Club! Join us and learn to cook some amazing meals! Every Wednesday after lessons, School Hall, 30 members, £5 a term</i></u> For B: Ask questions using the keywords below <i>When? Where? How many members? Cost?</i> (2 min)
2. Dinner (RP2) [A]	For A: You want to invite B to come to dinner at your house on Friday evening. If B agrees, talk about details. If B declines, you should negotiate with B about a possible date For B: You have another appointment for Friday evening. Say that you are sorry, that you will not be able to attend and the reason. If A suggests another date, agree if it is okay and talk about details (2 min)
3. Hobby (D1)	Have a conversation related to hobbies (e.g., <i>sports, clubs, last weekend, Golden Week</i>) (2 min)
4. Trip (D2) ^b	A and B have agreed to go on a trip together. Decide four things to bring for the trip, while asking each other questions <i>Place and time: Zoo in Hokkaido in January, Purpose: Seeing cute animals, Weather: Very cold at daytime and night</i> (2 min)
5. Job (RP3) ^c [A]	For A: <u>B is the owner of the shop where you are considering applying for a part-time job. Ask questions using the keywords below</u> <i>Name/shop? address? what/sell? telephone number? work every day?</i> For B: Use the information below and answer A's questions kindly <i>Happy Feet Store. We need a shop assistant to sell children's shoes. £6 per hour, Saturdays only: 9–5.30 pm, 8 Station Road, Phone 766814</i> (2 min)
6. Movie (RP4) [B]	For A: You are invited by B. Decline the offer at least once and explain the reason. Agree later if you like the suggested plan For B: You invite A for a movie. It can be viewed at a theater nearby and A will surely like it. If A agrees, talk about details. If A disagrees, convince A to see it on another day (2 min)
7. Friends (D4)	Have a conversation related to friends (e.g., <i>high school, university, part-time job, meet</i>) (2 min)
8. Date (D5)	Ken is A and B's friends. He is going on a first date with his girlfriend. He has asked A and B the following question. Discuss how A and B will give Ken advice <i>Question: Ken invited her for a date. Should he pay for everything? How much should he pay if he does not pay all?</i> (3 min)
9. Toothache (RP5) [A]	For A: <u>You have a toothache. Say how terribly it aches. Refuse B's advice at least once. Agree later and decide what to do if B suggests a plausible plan</u> For B: Show understanding of A's situation. Recommend that A go to the dentist and/or take a painkiller. Convince A by suggesting concrete plans (2 min)

(continued)

Table 1 (continued)

Task	Instruction
10. Driving (RP6) [B]	For A: Listen to B and tell B that you understand his/her feelings. Cheer B up. Give B advice on what to do next For B: You have failed a driving test. Tell A how sad you are. Decide what to do next based on A's advice (2 min)
11. Victory (RP7) [B]	For A: You won the game. B will celebrate the victory. Talk humbly. Say that you would like to appreciate those who helped you (anybody is okay) For B: Congratulate A, who won the game. Praise A even when A talks humbly (2 min)

Note () = task type No. [] = Who should begin the conversation. RP = Role play. D = Discussion. A/B = Students A/B. Underlined = information shared with A and B. The instructions were originally written in Japanese. Students were instructed to continue the related conversation after finishing the assigned task

^aDerived from Edwards (2008, p. 18–19)

^bAdapted from Butler and Zeng (2014)

^cDerived from University of Cambridge ESOL Examinations (2010, pp. 2–4)

Analyses

For rating, we used the same holistic rating scale of 1–3 in Koizumi et al. (in press), which considers interactional effectiveness and linguistic elements such as task achievement, fluency, accuracy, and appropriateness (see Table 2). We created a holistic scale since we weighed practicality over providing detailed feedback to students. We prioritized making a scale that enables teachers to evaluate by listening to the conversation once.

Using the scale, the three authors rated each talk independently. We had a 1-day rater training session assessing five pairs ($n = 10$), discussing any divergences, and

Table 2 Holistic rating scale

3	<i>Satisfies adequately</i> Satisfies the following task point(s). Communicates effectively in English by appropriately participating in turn-taking. Speaks fluently to the extent that the conversation is moving smoothly (Satisfies most of these abovementioned points.) E.g., Task 2 (Role-play 2: Dinner): The person who invites can do so appropriately and continue the related conversation The invited person can say no, apologize, and give reasons for not accepting the invitation appropriately and continue the related conversation
2	<i>Satisfies to a certain degree</i> Satisfies some of the task point(s). Communicates adequately in most everyday contexts but can be rather passive in responding and commenting (or mostly speaks alone, dominantly). Due to poor fluency, the conversation does not go smoothly, but the speaker aims to continue the conversation in English
1	<i>Needs more effort</i> Satisfies few task point(s). Gives simple responses only when required but is unable to maintain or develop the interaction. Stops the conversation unnaturally and does not make efforts to start it

adding some notes for the scale. We then evaluated the remaining students independently. One of the authors (Rater 1) rated all the remaining 180 students, Rater 2 rated 48, and Rater 3 rated 94. Scores from Raters 2 and 3 were combined and treated as Rating 2, while Rater 1's scores were considered Rating 1.

For MFRM, we used 190 students' scores, 11 tasks, and three raters, with missing values. For SEM and generalizability theory, we used 117 students' scores, 10 tasks, and two ratings, without missing values. The two groups can be considered similar because they had similar means and *SDs* of Rasch ability estimates ($M = 0.43$, $SD = 2.41$, $N = 190$; $M = 0.53$, $SD = 2.38$, $n = 117$). For SEM, we used a robust weighted least squares (WLSMV) estimation method and the software *Mplus* (Muthén and Muthén 2014) since the scores were on an ordered scale of 1–3 (RQ1). For MFRM, we used the rating scale model in an MFRM program, Facets (Linacre 2014; RQ2 to RQ5), to estimate the test takers' ability, task difficulty, rater severity, and rating scale. We performed generalizability theory using GENOVA (Center for Advanced Studies in Measurement and Assessment 2013) to calculate the number of tasks and raters needed to obtain highly consistent scores (RQ6).

Results and Discussion

Does the Test Have a Unitary Factor Structure Underlying the Paired Oral?

SEM allows us to construct models hypothesizing relationships between observed and latent variables, based on substantive theory and previous results, and to test whether these models fit the data well (see, e.g., Ockey and Choi 2015; Kline 2010). We hypothesized two models: a unitary model of one factor of oral interactive ability representing 11 tasks (Model 1) and an alternative model of two correlated factors (role-play and discussion abilities) representing two tasks each, as task formats may affect the structure (Model 2). For both models, we used Ratings 1 and 2 for each task (see Fig. 1).

Table 3 shows fit statistics for the unitary (Model 1) and correlated models (Model 2). Although the chi-square statistic was statistically significant ($\chi^2 = 371.921$, $df = 170$, $p < 0.01$; $\chi^2 = 371.612$, $df = 169$, $p < 0.01$) for both models, some indices showed a good fit (CFI = 0.95, TLI = 0.95), while others showed only a moderate fit (RMSEA = 0.10 [0.09, 0.12] and WRMR = 1.21). Model 2 was particularly problematic since its covariance matrix was not positive definite. One reason may be a correlation greater than or equal to 1 between two latent variables. The standardized path between the two-ability factors was 1.002. Model 2 was excluded from further consideration.

Model 1 was revised based on theory and modification indices. A revised model—Model 3—explained the data well (CFI = 0.97, TLI = 0.97, RMSEA = 0.08 [0.06, 0.10], WRMR = 0.99), with the parameter estimates presented in Table 4.

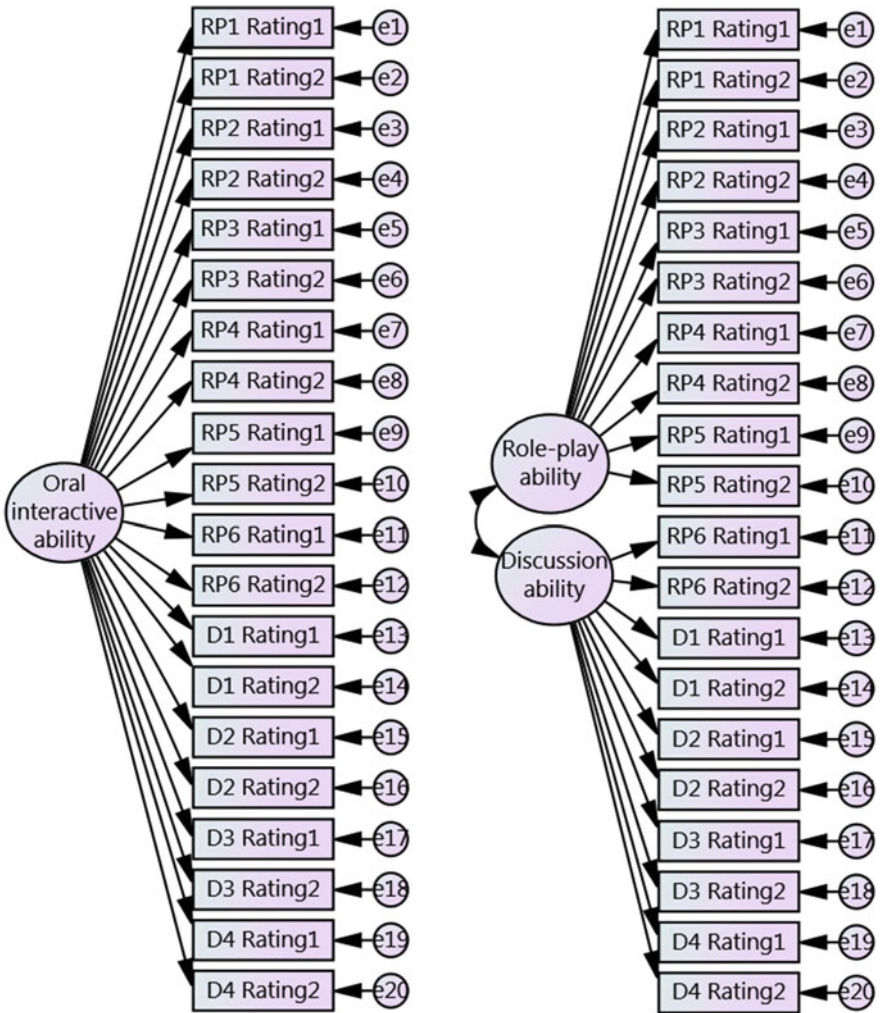


Fig. 1 Model 1 is on the *left* and Model 2 is on the *right*. Each observed variable is labeled by task type and rating. For example, RP1 Rating1 and D1 Rating1 refer to Role Play 1 Rating 1 and Discussion 1 Rating 1, respectively. See Table 1 for RP and D tasks

This suggests that the paired oral is considered to measure a single trait, which we interpret as oral interactive ability, which accords well with the intended test construct. The unitary structure adopted was the same overall as in Koizumi et al. (in press).

We also conducted MFRM and found that 45.05 % of the score variance was explained by Rasch measures, which also suggests unidimensionality of the structure. This percentage of the variance explained by Rasch measures was a little smaller than but similar to Koizumi et al. (in press; 57.90 %).

Table 3 Model fit indices

	χ^2 (df) <i>p</i>	CFI	TLI	RMSEA [90 %CI]	WRMR
Criteria	<i>p</i> > 0.05	>0.90	>0.90	<0.08	<1.00
Model 1: Unitary	371.921 (170) < 0.01	0.95	0.95	0.10 [0.09, 0.12]	1.21
Model 2: Two abilities correlated	371.612 (169) < 0.01	0.95	0.95	0.10 [0.09, 0.12]	1.21
Model 3: Unitary + correlated errors	283.184 (163) < 0.01	0.97	0.97	0.08 [0.06, 0.10]	0.99

Note *N* = 117. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval; WRMR = Weighted Root Mean Square Residual. The criteria were overall based on Byrne (2012) and Yu (2002, as cited in Wang and Wang 2012)

Table 4 Standardized parameter estimates for Model 3

	Path from oral interactive ability	Correlated error	Standard error	<i>R</i> ²
RP1 Rating1	0.75		0.05	0.56
RP1 Rating2	0.87		0.04	0.76
RP2 Rating1	0.76		0.05	0.58
RP2 Rating2	0.86		0.04	0.73
RP3 Rating1	0.54		0.08	0.29
RP3 Rating2	0.76		0.05	0.57
RP4 Rating1	0.73		0.06	0.53
RP4 Rating2	0.76		0.04	0.58
RP5 Rating1	0.74		0.05	0.54
RP5 Rating2	0.93		0.02	0.87
RP6 Rating1	0.75		0.06	0.56
RP6 Rating2	0.87		0.03	0.77
D1 Rating1	0.75		0.06	0.56
D1 Rating2	0.77		0.05	0.60
D2 Rating1	0.72		0.05	0.52
D2 Rating2	0.72		0.05	0.51
D3 Rating1	0.77		0.05	0.60
D3 Rating2	0.86		0.03	0.73
D4 Rating1	0.77		0.05	0.60
D4 Rating2	0.88		0.02	0.78
RP2R1 and RP1R1		0.71		
RP4R1 and RP3R1		0.49		

(continued)

Table 4 (continued)

	Path from oral interactive ability	Correlated error	Standard error	R ²
RP6R1 and RP5R1		0.63		
D2R2 and D2R1		0.55		
D2R2 and RP4R2		0.44		
RP4R2 and RP3R2		0.42		
D1R2 and RP2R2		0.54		

Note Each observed variable is labeled by task type and rating. For example, RP1 Rating1 and D1 Rating1 refer to Role Play 1 Rating 1 and Discussion 1 Rating 1, respectively. The path from RP1 Rating1 is set to 1 for identification. All other factor loadings are statistically significant

Do All Tasks and Raters Fit the Rasch Model?

Figure 2 displays the relationships between ability, task difficulty, rater severity, and rating scale. As seen in Table 5, test takers’ abilities spread very widely from -3.67 to 7.51. The task difficulty varied from -1.14 to 1.10, with Task 4 (Discussion 2: Trip) being the most difficult and Task 3 (Discussion 1: Hobby) the easiest. Koizumi et al. (in press) used similar tasks and the order was Task 1 (Role-play 1: Club, the most difficult), 4 (Discussion 2: Trip), 2 (Role-play 2: Dinner), and 3 (Discussion 1: Hobby). Compared to the current study, the order of difficulty of Tasks 4, 2, and 3 was the same. One reason Task 1 had a higher difficulty level than this study was that we added the context (e.g., *You are in a cooking club. B is considering whether to join it.*) of talking about a club; without the context, students must have found it hard to talk in the previous study. Because of the modification, the difficulty seems to have decreased at an appropriate level.

The rater severity differed across raters from -0.66 to 0.60, with Rater 2 as the most severe. Test-taker and task reliability were high (0.91–0.92), which shows consistency of scores across test takers and across tasks. High rater reliability (0.98) indicated that rater severity was different.

The infit mean square statistics between 0.5 and 1.5 were used to judge acceptable model fit (Linacre 2013). However, we did not regard an overfit as problematic (i.e., an infit mean square of below 0.5), because this indicates that the persons, tasks, and raters fit the model too well. We did not also regard an infit mean square between 1.5 and 2.0 as problematic, because it is “unproductive for construction of measurement, but not degrading” (Linacre 2013, p. 270). All the tasks and raters had values within this range, with 0.88–1.22 for the task and 0.83–1.10 for the raters. Furthermore, 15 students (7.89 %, 15/190) had values of less than 0.50 and were considered overfitting students, and 17 (8.95 %, 17/190) had infit mean squares of more than 1.5 and were considered underfitting students, but

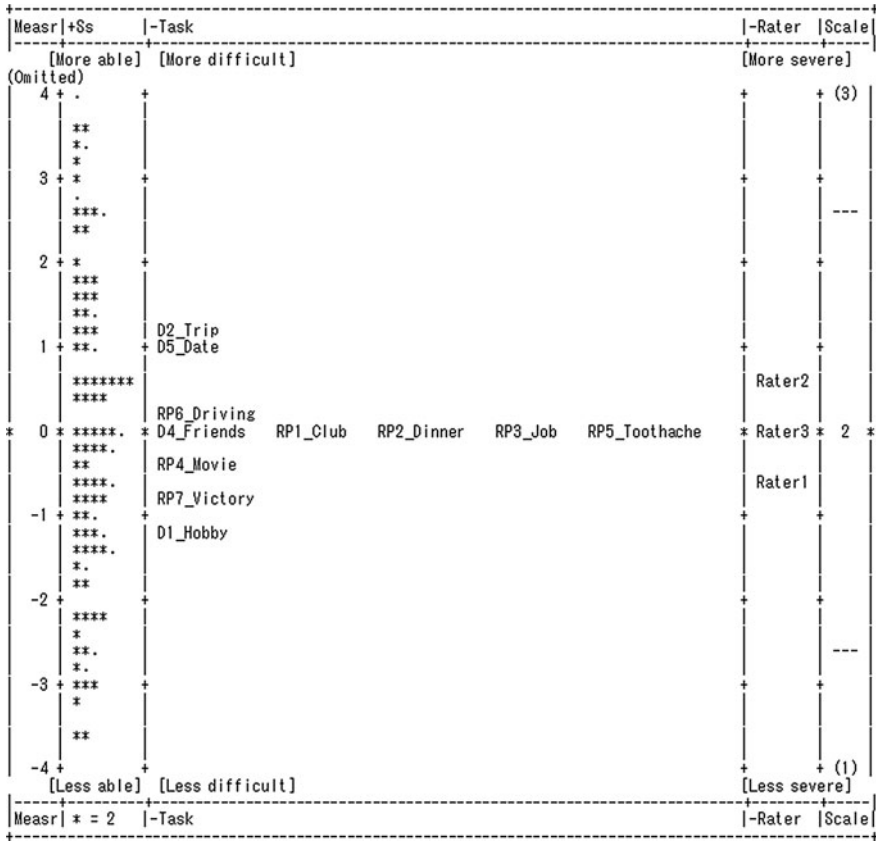


Fig. 2 Wright map for participants ($N = 190$), tasks ($k = 11$), raters ($n = 3$), and the rating scale. Ss = participants; * = 2 participants; . = 1 participant. Fifteen participants with measures of above 4.00 were omitted from the figure. Higher values mean higher ability in the second column, more difficult tasks in the third column, and more severe raters in the fourth column

Table 5 Descriptive statistics for the three facets

Test takers	Logit: $M = 0.43$, $SD = 2.41$; Min = -3.67 ; Max = 7.51 Fair average (on a scale of 1–3): $M = 2.07$; $SD = 0.43$; Min = 1.24 ; Max = 2.99 Reliability = 0.91 ; Separation = 3.11 ; Strata = 4.48 Infit mean squares: $M = 0.98$; $SD = 0.41$; Min = 0.04 ; Max = 2.13
Tasks	$M = 0.00$; $SD = 0.62$; Min = -1.14 ; Max = 1.10 Task reliability = 0.92 ; Separation = 3.34 ; Strata = 4.79 Infit mean squares: $M = 0.99$; $SD = 0.10$; Min = 0.88 ; Max = 1.22
Raters	$M = 0.00$; $SD = 0.52$; Min = -0.66 ; Max = 0.60 Rater reliability = 0.98 ; Separation = 6.34 ; Strata = 8.79 Interrater agreement = 69.6% ($1099/1580$); Expected agreement = 64.6% Infit mean squares: $M = 1.00$; $SD = 0.12$; Min = 0.83 ; Max = 1.10

Note The population SD s and reliability for the population with extremes are presented

one test taker had 2.13 and he somehow performed inconsistently across tasks, performing well in one difficult task and worse in an easy task. However, this was only one highly underfitting test taker. The fit of tasks and raters was all appropriate, which was in line with Koizumi et al. (in press).

Do Test Tasks Have a Wide Range of Difficulty and no Wide Gaps in Difficulty?

To see the distribution of tasks available in the test, we examined a task strata statistic, as Linacre (2013) recommends, because we statistically hypothesized that the measure distribution is not normal due to the many items at the peripheral end. We also expected that high and low difficulty levels in task measures are derived because of high and low task difficulty. The task strata was 4.79, which means 11 tasks could be classified into at least four different levels of task difficulty. Although this satisfied a minimum required level, we hope that we can differentiate each level into a few more, so higher task strata would be ideal.

Figure 2 shows that tasks were spread far less widely (range = 2.24, from -1.14 to 1.10) than test takers' abilities (range = 11.18, from -3.67 to 7.51) and that we should have more tasks at higher and lower ends of the scale, that is, more and less difficult tasks. Figure 2 also demonstrates the existence of some gaps on the logit scale. However, as seen in Table 6, most gaps were within the standard error of measurement and were regarded as not very substantive. For example, Tasks 11 (Role-play 7: Victory) and 3 (Discussion 1: Hobby) have a task difficulty of -0.77 and -1.14 respectively, but 68 % confidence intervals (CIs) overlapped (-1.18 to -0.36 and -1.28 to -1.00). There were two cases with different values beyond the standard error: between Tasks 8 (Discussion 5: Date) and 10 (Role-play 6: Driving), and between Tasks 2 (Role-play 2: Dinner) and 6 (Role-play 4: Movie). We can also argue that when we used 95 % CI, there was only a gap in the former case (0.71 to 1.21 and -0.01 to 0.49, not shown in Table 6 but calculated using $\text{Measure} \pm 1.96 * \text{SE}$), whereas there was an overlap in the latter (-0.34 to 0.20 and -0.63 to -0.13). Nevertheless, we decided to use 68 % CI to strictly improve our test. These two gaps in between as well as at the higher and lower ends can be modified in a future revision by adding tasks with such difficulty levels.

Is the Difficulty of Student Cards Equal?

As explained in the Method section, in the test, a student received a student card for Student A or B, and Students A and B made a pair. We compared the students' ability estimates across the two groups (Students A vs. B groups) but found no significant difference between the groups with the effect size being negligible

Table 6 Task measurement report

Task	Total count	Observed average	Fair average	Measure (logit)	Model SE	Infit MnSq	68 %CI (Measure ± SE)
D2: Trip	302	1.91	1.84	1.1	0.13	1.13	0.97 to 1.23
D5: Date	342	1.97	1.87	0.96	0.13	1.06	0.83 to 1.09
RP6: Driving	310	2.12	1.98	0.24	0.13	1.00	0.11 to 0.37
D4: Friends	339	2.12	2.00	0.10	0.13	0.95	-0.03 to 0.23
RP1: Club	282	2.13	2.01	0.03	0.14	0.93	-0.11 to 0.17
RP9: Toothache	335	2.15	2.01	-0.02	0.13	0.97	-0.15 to 0.11
RP3: Job	345	2.16	2.02	-0.05	0.13	1.22	-0.18 to 0.08
RP2: Dinner	280	2.15	2.02	-0.07	0.14	0.89	-0.21 to 0.07
RP4: Movie	346	2.21	2.07	-0.38	0.13	0.88	-0.51 to -0.25
RP11: Victory	39	2.62	2.13	-0.77	0.41	0.96	-1.18 to -0.36
D1: Hobby	314	2.36	2.19	-1.14	0.14	0.92	-1.28 to -1.00
Mean	294	2.17	2.01	0.00	0.16	0.99	-
SD ^a	83.90	0.18	0.01	0.62	0.08	0.10	-

Note SE = Standard error. MnSq = Mean squares. CI = Confidence interval

^aPopulation

(Student A: $M = 0.30$, $SD = 2.43$, $n = 95$; Student B: $M = 0.57$, $SD = 2.41$, $n = 95$; $t = -0.78$, $df = 187.99$, $p = 0.44$, $d = -0.11$, 95 % CI = -0.40 to 0.17). Thus, we can conclude that the difficulty level of student cards is considered equal. It should be noted that this result came from a situation where tasks requiring Student A to speak first are used almost the same number of times as tasks requiring Student B to do so; when teachers select tasks from the task pool, they may need to consider the balance of tasks from this perspective.

Does the Holistic Rating Scale Function Properly?

We analyzed functions of the rating scale based on Bond and Fox (2007). Table 7 indicated that results of the scale almost satisfied the criteria: There were more than 10 ratings at each level (420–1960). Thresholds, or difficulty estimates for choosing one level over another (e.g., -1.99 from levels 1–2) increased as the level increased, and the values of distances between thresholds between neighboring levels were 2.28 and 5.02; the former was between 1.4 and 5.0 logits but the latter was marginally beyond 5.0; we considered this to be minor. The probability curve (Fig. 3)

Table 7 Category statistics for the rating scale

Level	Number of observations (%)	Average measure for test takers at the level	Rasch-Andrich threshold measure (distance), standard error	Outfit mean squares
1	420 (14 %)	-1.99		1.0
2	1960 (65 %)	0.16	-2.51 (2.28), 0.06	1.1
3	645 (21 %)	2.62	2.51 (5.02), 0.06	1.0

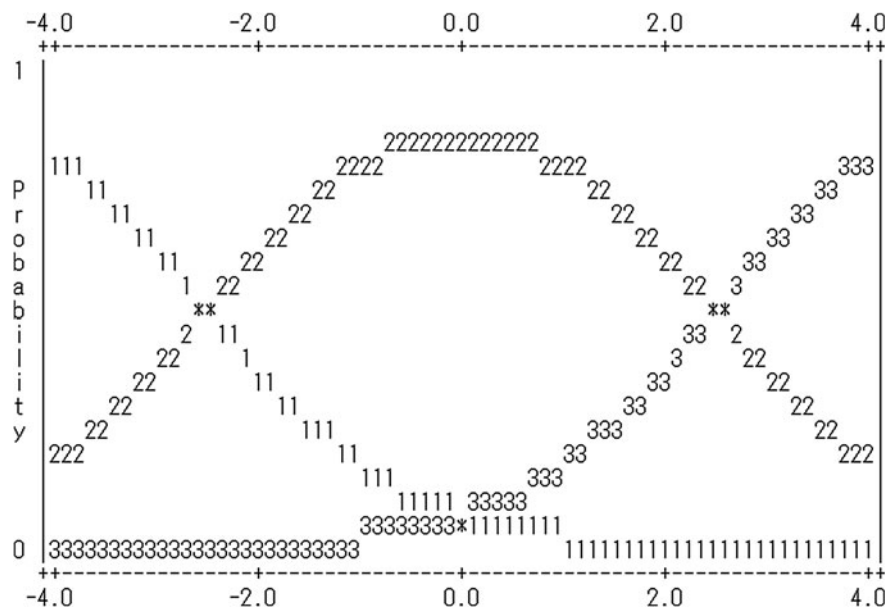


Fig. 3 The probability curve of the scale

had a clear top for Level 2. The level fit statistics were 1.0–1.1, less than 2.0. These results accorded well with the predicted patterns from the Rasch measurement and with Koizumi et al. (in press).

How Many Tasks and Raters Are Minimally Needed to Obtain Sufficient Reliability?

Using generalizability theory (Brennan 2001), we decomposed the test score variance into variance components affected by seven sources: variations of (a) persons’ ability (the objects of measurement), (b) task difficulty, (c) raters’ severity, (d) person-by-task interaction, (e) person-by-rater interaction, (f) task-by-rater interaction,

Table 8 Estimated variance component and percentage of variance explained

	<i>df</i>	Variance component	Percentage (%)	Standard error
Persons (<i>p</i>)	116	0.13	35.81	0.02
Tasks (<i>t</i>)	9	0.01	2.45	0.00
Raters (<i>r</i>)	1	0.02	4.34	0.01
<i>p</i> × <i>t</i>	1044	0.04	11.71	0.01
<i>p</i> × <i>r</i>	116	0.04	11.31	0.01
<i>t</i> × <i>r</i>	9	0.00	0.71	0.00
Residuals (<i>p</i> × <i>t</i> × <i>r</i> , <i>e</i>)	1044	0.12	33.67	0.01

and (g) the residual, consisting of the person-by-task-by-rater interaction and random errors, in the generalizability (G) study. We considered tasks and raters (which are actually ratings, with Rating 1 [scores from Rater 1] and Rating 2 [scores from Raters 2 and 3 combined]) as random facets. This method is often used for data in which not all raters evaluate all task responses (Lin 2014).

Table 8 shows the percentages of variance explained by the seven sources. The results suggest that the largest variability was explained by the persons (35.81 %), followed by the residual (33.67 %), and, to a lesser degree, by person-by-task (11.71 %) and person-by-rater (11.31 %) interactions. The percentages explained by tasks, raters, and task-by-rater interaction were marginal, ranging from 0.71 to 4.34 %. This suggests that the tasks and raters had similar levels of difficulty and severity. This appears in contrast to results from MFRM stating that task difficulty differed across tasks. However, MFRM results do not show the impact of tasks and raters on scores, and G study results showed that the impact was limited. The pattern in G study was almost the same as in Koizumi et al. (in press) except that the percentage of person-by-task interaction (11.71 %) was larger in the current study than in Koizumi et al. (5.79 %), probably because of an increased number of tasks.

Using the decision (D) study, we investigated how test reliabilities change depending on the number of tasks and raters. We used phi coefficients (Φ), which are used for an absolute decision, but results of generalizability (G) coefficients, for a relative decision, were also presented for interested readers. We employed a criterion of Φ = 0.70 or more, considering the use in low-stakes classroom assessment. Table 9 showed that when one rater evaluates the test, even the use of ten tasks does

Table 9 Phi coefficient (Φ) and generalizability coefficient (in the parenthesis) in decision studies (p × t × r design)

	1 task	2 tasks	3 tasks	4 tasks	5 tasks	6 tasks	7 tasks	8 tasks	9 tasks	10 tasks
1 rater	0.36 (0.39)	0.47 (0.51)	0.53 (0.58)	0.56 (0.61)	0.59 (0.64)	0.60 (0.65)	0.61 (0.67)	0.62 (0.68)	0.63 (0.69)	0.64 (0.69)
2 raters	0.48 (0.51)	0.60 (0.64)	0.66 (0.70)	<u>0.70</u> (0.74)	<u>0.72</u> (0.76)	<u>0.73</u> (0.77)	<u>0.74</u> (0.79)	<u>0.75</u> (0.80)	<u>0.76</u> (0.80)	<u>0.77</u> (0.81)
3 raters	0.54 (0.57)	0.67 (0.70)	<u>0.72</u> (0.76)	<u>0.75</u> (0.79)	<u>0.78</u> (0.81)	<u>0.79</u> (0.83)	<u>0.80</u> (0.84)	<u>0.81</u> (0.84)	<u>0.82</u> (0.85)	<u>0.82</u> (0.86)

Note Underlined = 0.70 or above

not lead to high reliability; when two raters join, at least four tasks are needed to obtain reliable scores; when three raters evaluate, at least three tasks are needed. In classroom assessment, usually one rater is available and in this case, a teacher may need to know that paired orals tend to have low reliability and to use as many tasks as possible. When two raters are available, the required number of tasks is reduced to four and this may be manageable. Koizumi et al. (in press) showed that conditions of one task with two raters, and three tasks with a single rater would produce sufficient reliability. This seems to indicate that when we increase tasks in the task bank, we should check the number of tasks and raters needed because this increase may change the impact of related factors on test scores.

Conclusion

We investigated six aspects related to the validity of the interpretation based on paired oral scores. We found that the structure of our paired oral has a unitary dimension, all tasks and raters fit the Rasch model, test tasks had a moderately wide difficulty range with gaps in between and at the higher and lower ends, the difficulty of student cards was equal, the holistic rating scale functioned properly, and the number of tasks and raters minimally needed to obtain sufficient reliability was at least four tasks with two raters and three tasks with three raters.

The results we obtained in this study were generally positive and as expected in the test developing stage. Major unexpected parts were the existence of gaps in between and at higher and lower ends of the scale, and they will be addressed and rectified in future research. We will also transcribe actual conversations and qualitatively examine relationships between linguistic functions intended to be elicited and those actually observed in the conversation. This information will help us identify what type of tasks should be included in the task bank together with the construct intended and the difficulty information that we obtained in the current study.

Our results will provide teachers with crucial information on how to use paired orals in their classroom. Moreover, we mainly used multifaceted Rasch measurement (MFRM), along with some auxiliary methods (structural equation modeling and generalizability theory) for the validation of our paired oral. MFRM has helped us identify strengths and weaknesses of our test and suggested improvements. The methods we used would be useful for other contexts where test takers, tasks, and raters are involved.

Acknowledgement This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C), Grant Number 26370737.

References

- Aryadousta, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly*, 13, 1–24. doi:10.1080/15434303.2015.1133626.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
- Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessment. *Language Assessment Quarterly*, 11, 45–75. doi:10.1080/15434303.2013.869814.
- Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.
- Cambridge ESOL Examinations (2010). *Speaking test preparation pack for Key English Test*. Cambridge, U.K.: Author.
- Center for Advanced Studies in Measurement and Assessment (University of Iowa, College of Education). (2013). *GENOVA suite programs*. Retrieved from <http://www.education.uiowa.edu/centers/casma/computer-programs#8f748e48-f88c-6551-b2b8-ff00000648cd>.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*.TM. New York, NY: Routledge.
- Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, 26, 367–396. doi:10.1177/0265532209104667.
- Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33, 117–135. doi:10.1177/0265532215582282.
- Edwards, L. (2008). *Common European Framework assessment tests*. London, U.K.: Mary Glasgow Magazines (Scholastic).
- Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, 5, 89–119. doi:10.1080/15434300801934702.
- Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, 35, 553–574. doi:10.1093/applint/amt017.
- Galaczi, E., & French, 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.
- Kley, K. (2015). *Interactional competence in paired speaking tests: Role of paired task and test-taker speaking ability in co-constructed discourse*. Unpublished Ph.D. dissertation, University of Iowa, U.S. Retrieved from <http://ir.uiowa.edu/etd/1663/>.
- Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Koizumi, R., In'nami, Y., & Fukazawa, M. (in press). Development of a paired oral test for Japanese university students. *British Council New Directions in Language Assessment: JASELE Journal Special Edition*.
- Lin, C.-K. (2014). *Treating either ratings or raters as a random facet in a performance-based language assessments: Does it matter?* CaMLA Working Papers 2014-01. Cambridge Michigan Language Assessments. Retrieved from <http://www.cambridgemichigan.org/sites/default/files/resources/workingpapers/CWP-2014-01.pdf>.
- Linacre, J. M. (2013). *A user's guide to FACETS: Rasch-model computer programs (Program manual 3.71.0)*. Retrieved from <http://www.winsteps.com/a/facets-manual.pdf>.
- Linacre, J. M. (2014). *Facets: Rasch-measurement computer program (Version 3.71.4)* [Computer software]. Chicago: MESA Press.
- McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, 29, 555–576. doi:10.1177/0265532211430367.

- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*, 241–256. doi:[10.1177/026553229601300302](https://doi.org/10.1177/026553229601300302).
- Muthén, L., & Muthén, B. (2014). *Mplus* (Version 7.2) [Computer software]. Los Angeles, CA: Muthén & Muthén.
- Negishi, J. (2015). Effects of test types and interlocutors' proficiency on oral performance assessment. *Annual Review of English Language Education in Japan*, *26*, 333–348.
- Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, *12*, 305–319. doi:[10.1080/15434303.2015.1050101](https://doi.org/10.1080/15434303.2015.1050101).
- Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing*, *26*, 325–339. doi:[10.1177/0265532209104665](https://doi.org/10.1177/0265532209104665).
- Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*, 411–440. doi:[10.1191/0265532206lt336oa](https://doi.org/10.1191/0265532206lt336oa).
- Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. West Sussex, UK: Wiley.