INDONESIAN JOURNAL OF APPLIED LINGUISTICS
Vol. 13 No. 2, September 2023, pp. 407-417

Available online at:
https://ejournal.upi.edu/index.php/IJAL/article/view/63075

https://doi.org/10.17509/ijal.v13i2.63075

On improving Kaiwa (会話) assessment: Incorporating JF standard descriptors and MFRM

Rina Supriatnaningsih[1*], Ahmad Yulianto[2], and Lispridona Diner[3]
[1,3]*Japanese Language Education Study Program, Faculty of Languages and Arts, Universitas Negeri Semarang, Kampus Unnes Sekaran Gunungpati 50229, Semarang, Central Java, Indonesia*
[2]*French Literature Study Program, Faculty of Languages and Arts, Universitas Negeri Semarang, Kampus Unnes Sekaran Gunungpati 50229, Semarang, Central Java, Indonesia*

ABSTRACT
Despite the available rubrics, assessing speaking objectively has been a debatable issue to language assessment experts mostly due to the dependence on the raters' authority. Scoring speaking performance often results in unfairness since subjectivity may come into play. Kaiwa (Speaking) is one of the four competencies examined in Japanese language assessment. On one hand, objective and accurate speaking assessment is badly needed. On the other hand, raters tend to overrate or underrate at times. Using Many-Facets Rasch Measurement (MFRM) and JF Standard descriptors, this study aimed to evaluate the Japanese speaking (Kaiwa) assessment. To this end, a cohort of 75 freshmen, consisting of 28 males (37%) and 47 females (63%), were assessed on the five-rubric scale (comprehension, vocabulary, structure, fluency, and pronunciation). These students' age ranged from 18 to 20 years of age and their Japanese proficiency level was equal to N5. Two raters were involved in the assessment. The result revealed that: (1) 29 biases were found in rater-student interaction and rater-component interaction; (2) different patterns of rating behaviour were discovered. Rater 1 was more lenient than rater 2 but rater 2 was more consistent; (3) pronunciation and fluency are components that contributed the most to bias while structure was the most objective component being scored. For examiners, this result implies that scoring moderation should be held before grading students. For policymakers, the implication of the study suggests that modifications in the assessment rubric and statistical control be made so that fairer ratings could be achieved.

Keywords: Assessment; bias; descriptor; Rasch, rubric

First Received:	**Revised:**	**Accepted:**
6 March 2023	19 June 2023	1 September 2023
Final Proof Received:		**Published:**
27 September 2023		30 September 2023

How to cite (in APA style):
Supriatnaningsih, R., Yulianto, A., & Diner, L. (2023). On improving Kaiwa (会話) assessment: Incorporating JF standard descriptors and MFRM. *Indonesian Journal of Applied Linguistics*, *13*(2), 407-417. https://doi.org/10.17509/ijal.v13i2.63075

INTRODUCTION

Speaking is viewed as an intangible construct that is not easy to measure reliably. This is associated with the dynamic and context-embedded nature of this competence and is also due to the different forms it assumes (e.g., monolog, paired conversation, or interview). Therefore, assessing speaking should be oriented to identify candidates' ability in using the target language orally (Noh & Matore, 2022). Fan and Yan (2020) reiterated the importance of rating quality produced by raters in speaking assessment. Each rater brings his/her own idiosyncrasies and values to the rating stage which causes scoring variability among raters (Engelhard et al., 2018). Grading norms are then very essential for speaking assessment (Polat, 2020; Seedhouse & Nakatsuhara, 2018).

The Japanese authority developed JF Standard by adopting the Common European Framework of Reference for languages. This framework offers guidelines to direct Japanese language teaching and learning, and evaluate the results. Nonetheless, it does not guarantee that the teaching, learning, and assessment of Japanese language proficiency run well and are free from problems. In Kaiwa or Japanese speaking assessment, raters are often

* Corresponding Author
Email: rinasupriatnaningsih@mail.unnes.ac.id

confronted with situations that lead to subjectivity as stated by Lee (2018). Since it engages two parties i.e., students and examiners, Kaiwa assessment should then take into account intervening factors that might affect the rating like test features, rating scales, and rater's effect. These elements make up the overall rating procedure in speaking assessment (Wind & Jones, 2019).

Many-Facets Rasch Measurement (MFRM) for Assessing Speaking

In speaking assessment, test-takers are usually appraised for fluency, accuracy, vocabulary, structure, and comprehensibility (Namaziandost, 2019). As a consequence, speaking evaluation should not only offer grading scales for giving certain marks but also describe pieces of evidence that can be presented to justify the grade accuracy (Sultana, 2019). Suzuki and Kormos (2020) contended that raters play an indispensable role in assessing examinees' speaking skills.

Lane (2019) maintained that accuracy and consistency of the rater in assessing an examinee's performance are of high importance. Rasch measurement approach views examinees and raters as different aspects or facets. As a probabilistic model, Rasch measurement is utilized for predicting the outcome of encounters between persons and assessment items (Linacre, 2018a). It also allows us to make probabilistic reports about item difficulty, candidate ability, and rating scales. Such statements are reported in terms of units called *logits*, the logarithm of the odds of a certain outcome (McNamara et al., 2019). The Many-Facets Rasch Measurement (MFRM) model was developed by Linacre (2019) to facilitate various facets or variables that generate an impact on the probability of individuals obtaining a given mark on test items from raters. The mathematical formula to measure the facets is given below:

$$\text{Log}\,(P_{nirk}/P_{nir}(k\text{-}1)) = B_n\text{-}D_i\text{-}T_r\text{-}F_k$$

where:

P_{nirk} = the probability of students *n* being rated *k* on item *i* by rater type *r*,
$P_{nir}(k\text{-}1)$ = the probability of students *n* being rated *k-1* on item *i* by rater type *r*,
B_n = the proficiency of student *n*,
D_i = the difficulty of component *i*,
T_r = the severity of rater type r, and
F_k = the difficulty scale of category k, relative to scale category k-1.

The underlying concept of Rasch measurement constitutes that discrepancy in test takers' performances is driven by a latent attribute or trait that the test expects to measure (Linacre, 2018b). MFRM offers a supreme tool to inspect raters' different characteristics such as consistency, leniency, and severity (Engelhard et al., 2018). Rasch measurement has contributed significantly to many areas of studies including speaking assessment (Aryadoust et al., 2019).

By incorporating the prevailing problems and current studies on Kaiwa assessment, the present study attempted to tackle these issues. First, most studies in speaking assessment only emphasized the importance of rubrics, descriptors, and rating scales without paying attention to the raters' bias and subjectivity. The current study aimed to overcome these predominant shortcomings. A second objective was to help examiners deal with different variables in speaking assessment measurement. Third, there are not many studies in Kaiwa assessment that estimate various components in relation to the accuracy of individual proficiency level. The present study was then to promote the effectiveness MFRM to deal with the raters' inconsistency Kaiwa assessment. MFRM turned out beneficial thanks to its multidimensional power i.e., ability to measure latent traits performed by test-takers or raters as well as test constructs and its delivery. All of these persist in the assessment of any language (Aryadoust et al., 2021).

In brief, this study aimed to address the following questions:

1) Do the scores of Kaiwa performance differ significantly across raters?
2) Is the Kaiwa grading across raters biased? If so, to what extent?
3) In what component is the Kaiwa grading most biased?

METHOD

The following steps were used in this study: (1) setting an appropriate task to elicit a representative sample of the population; (2) ensuring the validity and reliability of elicited sample and its scoring. These two steps were of great importance for an appropriate task would enable us to obtain the data which, otherwise, would not be possible. The purpose of establishing validity and reliability was essential to ensure that data obtained were sound and replicable, and the results were accurate. With this, the integrity and quality of a measurement instrument were guaranteed. Adjustment was made following the need for Kaiwa assessment. As interview was used as a tool for obtaining the data, the appropriateness of the interview was first

assessed, then followed by the inspection of its validity and reliability.

Participants

Participating in this study were 2 female Japanese lecturers who hold doctoral degrees in language education and have taught Japanese for 16 - 35 years (M = 25.5, SD = 13.44) and 75 freshmen. They were enrolled in the Japanese Department of Universitas Negeri Semarang. They were made up of 28 male students (37 %) and 47 female students (63 %). Their age ranged from 18 to 20 years. All participants had received 1 year of Japanese language instruction by the time of the study. They were chosen for this study considering that examiners were often generous to freshmen.

Table 1

Rater Profile

	Academic Qualification	Japanese Competence	Teaching Experience	Expertise in Kaiwa rating	Familiarity with JF Can-dos
Rater 1	Doctoral Degree	Equal to N1	35 years	Advanced	Advanced
Rater 2	Doctoral Degree	Equal to N2	16 years	Moderate	Moderate

Instrumentation

As instruments, this study used Kaiwa test and JF Can-dos rubrics to score the participants' performance. The objective of Kaiwa test was to measure students' ability in producing meaningful, coherent, structurally correct, and appropriate sentences. The test was conducted in 2 stages i.e., brief presentation on a chosen topic, and an interview with the examiner. The scoring was done using rubric and rating scales which focus on five components namely comprehension (理解), vocabulary (語彙), structure (文法), fluency (流暢さ), and pronunciation (発音). The rubric contained 5 bands ranging from poor (ぜんぜんできない), bad (あまりできない), fair (まあまあ), good (できた), and excellent (素晴らしい). The examiner gave a mark for each criterion, choosing from a range of marks (1-5) according to the student's performance. Each criterion is equipped with a set of descriptors. The topic was selected from JF Can-dos and daily activities to minimalize rater's preference (Polat, 2020). In the interview, both examiners attempted to use more or less the same expressions.

Procedures

Before attending the test, all participants were given clear instructions and told that their identities were kept confidential. The present study employed a fully-crossed design where all raters rated all participants. In the test, the participants were required to choose a simple topic to talk about. Afterward, they presented the topic for 1 to 2 minutes. And then, the examiner asked them questions about what they had presented in the monologue. This interview lasted for 10 minutes. To minimize the fatigue effect both on the examiners and the students, each session of the test was limited to a maximum of 20 participants.

The use of 2 examiners for each test-taker was preferred in search of inter-reliability of the raters and to get a common perspective on the judgment. The raters assessed the Kaiwa proficiency using the descriptor and rating scales/bands. The obtained scores were subjected to descriptive statistics and MFRM analyses done on SPSS 25 and Facets 3.8.

Data collection and analysis

The data were collected from the scores given by the raters for the participants' performance. They were checked on 5 components namely comprehension, vocabulary, structure, fluency, and pronunciation. To ascertain that the data were valid and could achieve the aim of the study, the reliability of the rating scale and Rasch assumptions had been met beforehand. Cronbach's Alpha showed a strong correlation coefficient value of .89 for rating scale reliability. Therefore, the data were valid, reliable, and suitable for this study.

Linacre (2022) stated that Rasch analysis has three assumptions i.e., unidimensionality, local independence, and model-data fit. Since these three assumptions are related to one another, then satisfying one assumption guarantees the others; and testing the model-data fit is the most appropriate step to undertake. This assumption was verified by examining standardized residuals. The number of standardized residuals outside ± 2 range should not exceed 5% of the total data, and those outside ± 3 should not exceed 1% (Linacre, 2018b). The total data computed in this study was 750, involving 75 students, 5 components, and 2 raters (75x5x2). One (0.13%) standardized residual outside ± 2 and one (0.13%) standardized residual outside ± 3 were found. It indicated an adequate model-data fit, and that the assumptions of Rasch analysis had been satisfied.

Data analysis was run in two stages. In the first stage, the raw score was analysed on descriptive statistics. This step was important to get a preliminary description of the 2 raters' scoring tendency. In the second stage, Rasch analysis was performed to find out if there were biases. Three facets specified were students, raters, and test components.

FINDINGS AND DISCUSSION

RQ1. Do the scores of Kaiwa's performance differ significantly across raters?

To respond to this question, the result of descriptive statistics analysis was checked to find out whether significantly different scores occurred between the 2 raters. After that, Rasch analysis was run to get a more detailed pattern of the 2 raters' differences or similarities. Table 2 shows that the comprehension score given by rater 1 to students (*M* = 4.16, *SD* = .64) was higher than that of rater 2 (*M* = 3.84, *SD* = .37). A comparison of vocabulary also demonstrates that rater 1 score (*M* = 3.95, *SD* = .70) was slightly higher than rater 2 (*M* = 3.87, *SD* = .41). In structure comparison, rater 1 also gave a higher score (*M* = 4.08, *SD* = .63) than rater 2 (*M* = 3.95, *SD* = .32). Fluency comparison shows that rater 1 gave higher score (*M* = 4.28, *SD* = .61) than rater 2 (*M* = 3.95, *SD* = .32). Pronunciation is the only component where rater 1 (*M* = 3.48, *SD* = .60) gave lower score than rater 2 (*M* = 3.68, *SD* = .47).

Table 2

Mean Divergence across Raters

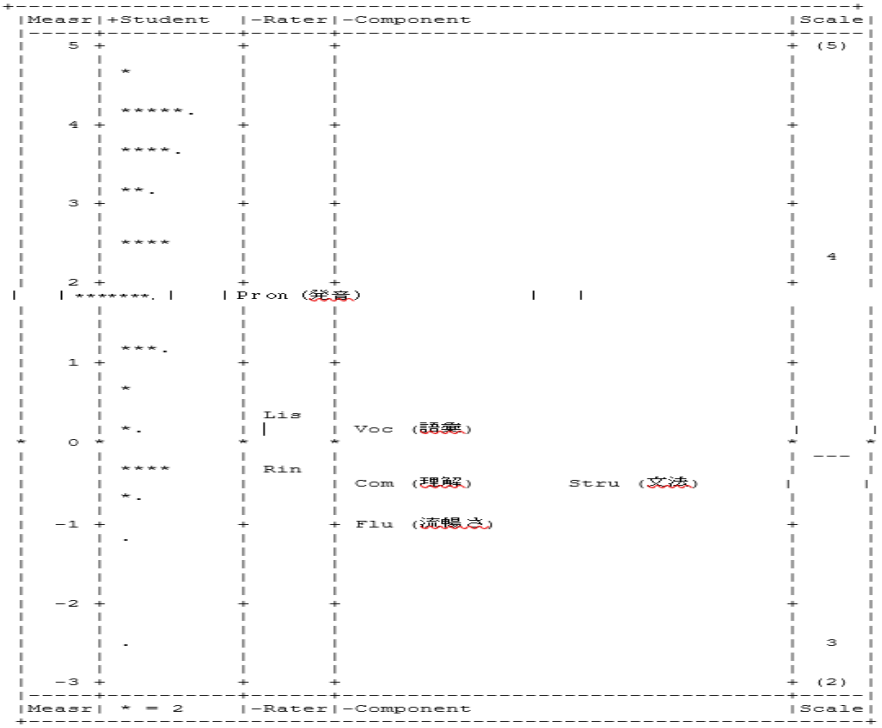
	Rater	*M*	*SD*	Min	Max	*N*
Com (理解)	1	4.16	.64	3	5	75
	2	3.84	.37	3	4	75
Voc (語彙)	1	3.95	.70	3	5	75
	2	3.87	.41	3	5	75
Struc (文法)	1	4.08	.63	3	5	75
	2	3.95	.32	3	5	75
Flu (流暢さ)	1	4.28	.61	3	5	75
	2	3.95	.32	3	5	75
Pron (発音)	1	3.48	.60	2	4	75
	2	3.68	.47	3	4	75

**p<.05*

Figure 1 presents an overall view of the interaction among students, rater, and components. Three facets i.e., students' overall competence, raters' inconsistency, and test components were measured. Column 1 shows the measurement unit (*Measr*) presented in *logits*. Column 2 shows the dispersion of students represented in asterisk (*) and dot (.). Column 3 shows the raters and their grading severity; the higher the more severe and the lower the more lenient. Column 4 shows Kaiwa test components. Column 5 shows the rating scale used in the grading, spanning from 1 to 5 where 1 is the lowest and 5 is the highest.

Figure 1

All Facet Vertical Rulers



Note: one asterisk (*) indicates 2 ratees while one dot (.) 1 ratee.

As displayed in column 1, there was a large distance in students' ability from -3 to +5 logits. Column 2 indicates how students' abilities are clustered. Two students are positioned slightly below +5 logits. Eleven students are found above +4 logits. Nine are found slightly below +4, and 5 students are slightly above +3. Eight students are found between +2 and +3 logits. The biggest group of 15 students is placed a little bit below +2 logits. Seven candidates are a bit above +1. Five students are spread between +1 and 0. Eleven students are distributed between 0 and -1. One is located slightly below -1 and one is located in the middle between -2 and -3. Column 3 presents the rank of raters from severe to lenient as represented by its span from +1 to 0 logits, suggesting that the 2 raters are not much different in grading the students. Rater 2 was more stringent than rater 1. Column 4 describes Kaiwa test components namely comprehension, vocabulary, structure, fluency, and pronunciation. For the students, pronunciation was the hardest component to achieve with the logit value slightly below 2. Vocabulary seems to be the fairest component in the assessment as shown by its logit value close to 0. Comprehension and structure have the same logit value between 0 and -1, suggesting that these components are relatively easy for the students. Found exactly at -1 logits value, fluency appears to be the easiest component to achieve. As for the scale used in the Kaiwa grading, the points extend from 2 to 5 indicating that no students got 1 point.

Table 3
Rater Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit		Outfit		Rater
						MnSq	ZStd	MnSq	ZStd	
1496	375	3.99	4.00	-.35	.12	1.16	2.0	1.12	1.2	1 (Rin)
1446	375	3.86	3.91	.35	.12	.84	-2.0	.84	-1.7	2 (Lis)
1471.0	375	3.92	3.95	.00	.12	1.00	.0	.98	-.2	Mean (Count:2)
25.0	.0	.07	.05	.35	.00	.16	2.1	.14	1.5	S.D. (Population)
35.4	.0	.09	.06	.50	.00	.22	2.9	.20	2.2	S.D. (Sample)

Model, Populn: RMSE .12; Adj (True) S.D. .33; Separation 2.79; Reliability .89
Model, Fixed (all same) chi-squared: 17.6; d.f.: 1; significance (probability): .00
Inter-Rater agreement opportunities: 375; Exact agreements: 191 = 50.9%; Expected: 235.9 = 62.9%

Table 3 shows that rater 1 who had assigned 3.99 points (Obsvd Average) was more lenient than rater 2 who had given 3.86 on average. This point average represents all Kaiwa 5 components listed in the rubric. The RMSE (Root Mean Square Standard Error) value indicates the standard error mean value for the total data except for the outliers. Its value of .12 points suggests that the standard error mean was noticeably low in this study. To verify the RMSE value, the Adjusted Standard Deviation was also checked, and was found to be .33. This is well below the critical level of 1.0. This result was considered reliable since it was higher than the agreed-upon value of .85 (Linacre, 2018b). With the infit (1.16) and outfit (1.12) of rater 1 and (.84) and (.84) of rater 2 well below the critical limit value of 1.4 logits, it leads us to believe that the raters' judgment is still reliable and acceptable. None of the two raters were identified as misfitting. They were relatively consistent in scoring.

Albeit the difference in severity level between rater 1 and rater 1 was only .70 logits (-.35 to +.35), the hypothesis that "*there is no statistical difference in the raters' scoring*" was rejected provided that discrepancies occurred between the 2 raters ($\chi 2$ = 17.6, df = 1, $p < .05$). The reliability value of .89 point and separation index of 2.79 which were quite high indicate the likelihood of the two raters to be consistently different in terms of stringency. Rater 1 was likely to be more lenient than rater 2; conversely, rater 2 was likely to be harsher than rater 1. It is further confirmed by the inter-rater exact agreement value of 191 out of 375 opportunity values, meaning that only 50.9% was achieved. This is quite far from the expected agreement value which should be at least 235.9 or 62.9%.

The statistical analyses indicate that there were differences and similarities in scoring between rater 1 and rater 2. Descriptive statistics analysis shows that the 2 raters showed divergent patterns in judging the students' performance. Rater 1 labelled the students as good in 3 (comprehension, structure, and fluency) out of 5 components as confirmed by *M* average which was higher than 4. Meanwhile, none of the means of scores for 5 components given by rater 2 exceeded fair predicate as confirmed by *M* average (3.68-3.95). Second, to these 2 raters, fluency was the most noticeable aspect in Kaiwa since both gave it the highest score. This is in line with JF Can-dos paradigm (2010: 79) where for spoken competence at a lower level, the ultimate goal to achieve is always marked with "I can". At A1 and A2 levels, fluency is emphasized, unlike B1 where coherence and cohesion are also taken into consideration. It conforms Suzuki and Kormos's account (2020) that for fluency at lower level what matters is as long as the speaker sounds smooth in the listener's point of view since it guarantees the success of communication. Both raters were aware of this premise considering their teaching experiences and familiarity with JF Can-dos. Interestingly, the 2 raters did not treat vocabulary

scoring as set out in the JF Can-dos where the ability for vocabulary is not too demanding …. *Have a basic vocabulary repertoire of isolated words and phrases related to particular concrete situation* … As such, the 2 raters should not have been too severe, contrary to what they did instead. Pronunciation was the component to which rater 1 and 2 gave the lowest score.

In principle, the scores from rater 1 and rater 2 are not significantly different. As displayed in Figure 1 and Table 2, each rater was consistent in scoring where rater 1 was likely lenient while rater 2 was likely severe. However, further investigation shows that rater 2 demonstrated a higher degree of consistency in scoring than rater 1. This is in contrast with Fan and Yan (2020) who argued that raters' features such as language background, experiences, and qualifications may have considerable impact on rating behaviour. The more experienced a rater and the more training s/he has had, the more impartial and consistent s/he is in rating. Rater 2 who has got fewer years of experience and training than rater 1 turned out to be more impartial and dependable. Yet, it is just in line with Koizumi and Watanabe (2021) who stated that raters are varied in their level of stringency and consistency. Some raters are more stringent while others are more generous. Some are steady in stringency throughout the grading process, and across rubrics; others aren't, unconsciously readjusting rating styles and severity levels. Rater 1 shows some inconsistencies concerning students' performance scores across components. Rater 2 demonstrates consistency both to students' performances and Kaiwa components.

RQ2. Is the Kaiwa grading across raters biased? If so, to what extent?

To answer this research question, bias interaction report between rater and student was analyzed and interpreted as shown in table 4.

Table 4

Bias Interaction between Rater and Student

Student No	Measr (Ability)	Rater	Obsvd Score	Expected Score	Obs-Exp Average	Bias (logits)	S.E.	*t*	Infit MnSq
7	1.83	Rin	24	19.77	.85	4.57	1.24	3.68	.4
67	1.83	Lis	23	19.22	.76	4.06	1.00	4.04	.8
46	1.22	Rin	23	19.30	.74	3.97	1.00	3.95	.8
20	-.27	Lis	20	17.09	.58	3.11	1.14	2.73	.1
23	-.27	Lis	20	17.09	.58	3.11	1.14	2.73	.1
26	-.27	Lis	20	17.09	.58	3.11	1.14	2.73	.1
58	-.27	Lis	20	17.09	.58	3.11	1.14	2.73	.1
28	3.09	Lis	23	20.19	.56	2.80	1.00	2.79	.8
9	3.64	Rin	24	21.33	.53	2.76	1.24	2.23	.4
39	.18	Lis	20	17.62	.48	2.66	1.14	2.33	.1
70	.18	Lis	20	17.62	.48	2.66	1.14	2.33	.1
2	-.27	Rin	20	17.91	.42	2.41	1.14	2.11	1.9
41	-.27	Rin	20	17.91	.42	2.41	1.14	2.11	.1
43	3.09	Rin	23	20.79	.44	2.10	1.00	2.09	.6
74	3.09	Rin	23	20.79	.44	2.10	1.00	2.09	.6
48	2.47	Lis	18	19.72	-.34	-1.95	.96	-2.03	.5
39	.18	Rin	16	18.38	-.48	-2.08	.94	-.2.22	.5
70	.18	Rin	16	18.38	-.48	-2.08	.94	-2.22	.7
43	3.09	Lis	18	20.19	-.44	-2.57	.96	-2.66	.5
74	3.09	Lis	18	20.19	-.44	-2.57	.96	-2.66	.7
20	-.27	Rin	15	17.91	-.58	-2.58	1.00	-2.57	1.1
23	-.27	Rin	15	17.91	-.58	-2.58	1.00	-2.57	1.1
26	-.27	Rin	15	17.91	-.58	-2.58	1.00	-2.57	.2
58	-.27	Rin	15	17.91	-.58	-2.58	1.00	-2.57	1.1
9	3.64	Lis	18	20.65	-.53	-3.12	.96	-3.23	.5
28	3.09	Rin	18	20.65	-.56	-3.27	.96	-3.39	.5
46	1.22	Lis	15	18.70	-.74	-3.36	1.00	-3.35	.2
67	1.83	Rin	16	19.77	-.75	-3.74	.94	-3.98	.7
7	1.83	Lis	15	19.22	-.84	-3.97	1.00	-3.96	.2

Note: This table displays significant biases only where t score fell below -2.0 or above +2.0.

Column 1 displays students' number. Column 2 indicates their ability measured. Column 3 presents the rater's identity. Column 4 shows the observed scores. Column 5 presents the expected score. Column 6 displays the margin between the observed score and the expected score (the value can be plus or minus). Column 7 shows bias logits which represent the different level and the likely error of the bias estimate. Column 8 presents standard error. Column 9 displays bias estimate converted into *t* score for each student. Column 10 presents infit mean square value. There are 29

significant biases involving 16 students. There are thirteen students experienced double biases by rater 1 and rater 2 namely students numbered 7, 9, 20, 23, 26, 28, 39, 43, 46, 58, 67, 70, and 74. Out of 29 biases, only one was detected as misfitting where its fit value was 1.9 above the tolerated 1.5 logits.

MFRM analyses detected rater misbehaviour which in this study was limited to leniency against severity. Pertinent in this discussion are raters' difference judgment of performance and degree of severity. The result shows that rater 1 and rater 2 made biased judgement with 15 and 14 conducts respectively. Rater 1 demonstrated leniency meanwhile rater 2 severity. Take, for instance, student no.7 (whose ability estimate was 1.83 logits) obtained a total score of 24 for five components while in fact his expected score was only 19.77. It suggests that the score given by rater 1 was higher or more lenient than it should have been. There was a .85 margin; deducted from (24-19.77)/5=0.85. Student no.7 was also disadvantaged from score bias by rater 2; 15 on average down 4.22 point from his expected score of 19.22. Unlike the score from rater 1, this score was just too stringent with -.84 margin (15-19.22)/5= -.0.84.

The 2 raters' score awarded to student no.7 is an example of how raters can be similar and different at the same time. They are the same in the sense that both raters' gradings are biased but how they are biased is different. As experienced lecturers, both raters have undertaken some training. Therefore, they were supposed to be fair in judging the students. It is worth discussing the 2 raters' misbehaviors which are not much different as confirmed by the mere gap between rater 1's point margin of .85 and rater 2's point margin of .84. Rater 1 who is more experienced apparently overrated student no.7 while rater 2 who is less experienced underrated him. In fact, both raters had undergone some training in Japanese language assessment.

Table 4 indicates how the 2 raters' biases are contrastive to one another. What was perceived as good by rater 1, was regarded otherwise by rater 2. This circumstance was confirmed by the inter-rater agreement which was only 191 (50.9%), quite far from the expected probability of 235.9 (62.9%) as shown in Table 3. However, Bond et al., (2021) maintained that "*as long as a rater is internally consistent in the way s/he applies the rubric, it does not matter even if difference in severity occurs*." The raters' biases, be they severity or leniency, are not closely related to students' ability. For instance, student no.9 who had the highest ability index (3.64) obtained a bias value of 2.76 from rater 1 and -3.12 from rater 2. Students no. 39 and 70 who came up with the least ability index of .18 received bias value of -2.08 from rater 1 and 2.66 from rater 2. The greatest bias from rater 1 occurred to student no.7 (4.57) while the greatest bias from rater 2 occurred to student no.67 (4.06). Having ability index of 1.83, these 2 students belong to midpoint groups. So, it indicates that the raters' bias happened to any students regardless of their ability. Some less competent students were overrated but some others with approximately similar competencies suffered from strict judgment. It indicates that individual preferences or intimacy between rater and students did not become influential factors in the rating.

Bias is detrimental to students since it might discourage them, reduce their motivation, and make them overestimate their ability. Students with low-level ability but receiving higher grades from lenient teachers will not put the effort in their endeavour in the future. They think they are good enough. On the other hand, students with high level ability but receiving lower mark will be discouraged and might think what they have done is useless. As for teachers, mistake in rating the students can downgrade their professionalism (Peabody & Wind, 2019) because psychometrically sound marks are not given to students when they are too stringent or generous.

Figure 2

Bias Interaction Graph between Rater and Student

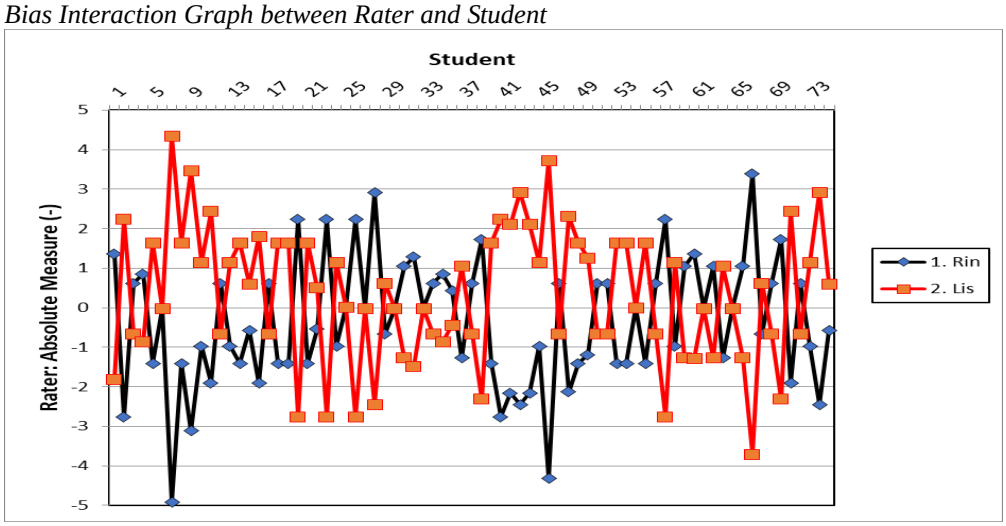


Figure 2 displays plots of rater-student interactions where raters are plotted on the vertical axis and students are presented on the horizontal axis. The black line represents rater 1 while red line represents rater 2. It further indicates that there are bias patterns pertinent to each rater. Rater 1 patterns tend to show negative logit values whereas rater 2 patterns lean towards positive logit values. This reiterates what has been discovered earlier that rater 1 is somewhat lenient while rater 2 is rather severe in the scoring.

Concerning the interaction between severity level and raters' background, no fixed pattern was found. This denies what was proposed by Weilie (2018) that severity level decreases as teaching experience increases. Both raters demonstrate an unpredictable pattern in terms of rater-to-student interaction. Noh and Matore (2022) argue that rating quality constitutes a fundamental property of effective and justifiable assessment. The analyses of the present study suggest that the 2 raters do not fully meet the quality rating which in turn damages the students. They failed to see the assessed constructs resulting in a wrong interpretation of the candidates' actual capability. If we observe the figure above, it is worth investigating further how the raters' tendencies are clustered in an unorganized manner. All and all, it should be highlighted that although both raters are experienced, of similar qualifications, working in the same department and collaboratively constructing the assessment, there was still a difference in interpreting the rubric leading to different severity.

RQ3. In what Kaiwa component is the grading most biased?

Table 5 below presented how each component contributed to the assessment bias. Table 5 shows that fluency is the component to which the raters gave the most generous score (4.11), higher than structure (4.01), comprehension (4.00), vocabulary (3.91), and pronunciation (3.58). Column 5 (Measure) displays the difficulty level down the way from pronunciation with the highest logits (1.84), followed by vocabulary (.09), comprehension (-.42), structure (-.49), and at last fluency with the lowest logit value (-1.02). So, it was difficult for the test-takers to obtain higher scores in pronunciation compared to the other 4 components while it was relatively easier for them to gain better scores in fluency. It spanned from +1.84 (the toughest) to -1.02 (the easiest) or 2.86 gap. Above of all, the subtle difference of *Obsvd Average* and *Fair (M) Average* indicates that the score given for each component was quite fair. Although RMSE (*Root Mean Square Standard Error*) value was relatively small (.19), but the *Adjusted Standard Deviation* value (.97), separation index (5.13), and reliability (.96) were somewhat high. This reveals that no statistically equal scores were given across these 5 components. The infit values suggest that no component was identified as misfitting and that bias patterns among components were consistent.

Table 5

Component Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit		Outfit		N Component
						MnSq	ZStd	MnSq	ZStd	
537	150	3.58	3.61	1.84	.19	1.10	.8	1.03	.2	5 Pron (発音)
586	150	3.91	3.94	.09	.19	1.22	1.6	1.24	1.5	2 Voc (語彙)
600	150	4.00	4.00	-.42	.19	.97	-.2	.93	-.4	1 Com (理解)
602	150	4.01	4.01	-.49	.19	.90	-.8	.88	-.8	3 Stru (文法)
617	150	4.11	4.08	-1.02	.19	.83	-1.6	.83	-1.2	4 Flu (流暢さ)
588.4	150.0	3.92	3.93	.00	.19	1.00	.0	.98	-.1	Mean (Count: 5)
27.5	.0	.18	.16	.98	.00	.14	1.2	.14	1.0	S.D. (Population)
30.8	.0	.16	.18	1.10	.00	.16	1.3	16	1.1	S.D. (Sample)

Model, Populn: RMSE .19; Adj (True) S.D. .97; Separation 5.13; Reliability .96
Model, Fixed (all same) chi-squared: 140.0; d.f.: 4; significance (probability): .00
Model, Random (normal) chi-squared: 3.9; d.f.: 3; significance (probability): .27

Table 6 below shows that the greatest bias was done by rater 2 with a logit of .87 for pronunciation. She overrated the students by .17 point on average. The second greatest bias was done by rater 1 with a logit of .53 point for fluency. She overrated the students by .10 point on average. The lowest bias was done by rater 1 with a logit of -.86 for pronunciation. She underrated the students down by -.10 point on average. Overall, there are 10 bias interactions between rater-to-components and both raters contributed the same frequency with 5 biases by each of them. Biases occurred to all components. It is interesting to note that pronunciation turned out to be the greatest bias (.87) as well as the lowest bias (.86) done by the raters. The graph below illustrates the bias interaction between rater and component.

Table 6

Bias Interaction between Rater and Components

Obsvd Score	Expctd Score	Obs-Exp Average	Bias	Model S.E.	*t*	Infit MnSq	Outfit MnSq	Rater	Component	Measr
276	263.36	.17	.87	.27	3.26	1.1	1.1	Lis	Pron (発音)	1.84
321	313.44	.10	.53	.26	1.99	.9	.9	Rin	Flu (流暢さ)	-1.02
312	304.85	.10	.50	.26	1.90	1.2	1.2	Rin	Com (理解)	-.42
290	288.06	.03	.14	.27	.53	1.0	1.0	Lis	Voc (語彙)	.09
306	305.86	.00	.01	.26	.04	1.1	1.1	Rin	Stru (文法)	-.49
296	295.94	.00	00	.27	.02	.7	.7	Lis	Stru (文法)	-.49
296	297.76	-.02	-.13	.27	-.47	1.5	1.5	Rin	Voc (語彙)	.09
288	294.95	-.09	-.51	.27	-1.87	.6	.6	Lis	Com (理解)	-.42
296	303.34	-.10	-.52	.27	-1.95	.7	.7	Lis	Flu (流暢さ)	-1.02
261	273.65	-.10	-.86	.26	-3.36	.7	.7	Rin	Pron (発音)	1.84
294.2	294.12	75.0	.00	.27	.01	.9	.9	Mean (Count: 10)		
16.4	14.60	.0	.51	.00	1.93	.3	.3	S.D. (Population)		
17.2	15.39	.0	.54	.00	2.04	.3	.3	S.D. (Sample)		

Note: Fixed (all = 0) chi-squared: 37.3; d.f.: 10; significance (probability): .00

Figure 3

Bias Interaction between Rater and Component

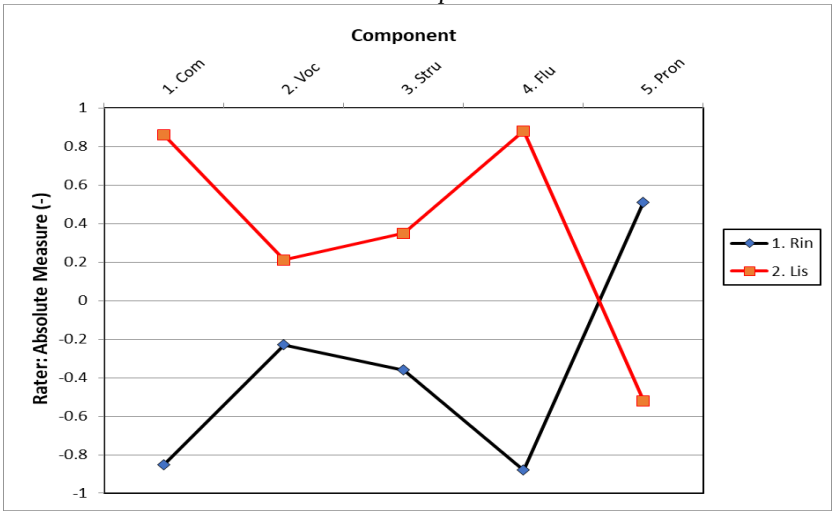


Figure 3 graphically plots rater-component interactions where raters are plotted on the vertical axis and components are displayed on the horizontal axis. Black line represents rater 1 while red line represents rater 2. It shows that there are bias patterns pertinent to each rater with respect to components tested. Rater 1 patterns show negative logits value almost for all components except pronunciation. Her leniency levels are described as follows. She is somewhat lenient when giving score for comprehension (-0.8), then getting closer to objective for vocabulary (-0.2), becoming more lenient for structure (-0.28), coming back to leniency for fluency (-0.9), but then being somewhat stringent for pronunciation (0.6).

Whereas rater 1 patterns are likely to be below 0 and are inclined to leniency, rater 2 patterns are on the contrary as leaning towards positive logits value. She was quite strict for comprehension and fluency with logit values of nearly 0.9. The score she gave for vocabulary (0.2) and structure (0.36) were close to the objective point. However, her logit value dropped for pronunciation (nearly -0.5), indicating her turn to leniency.

Concerning rater-to-component interaction, the finding of the present study is consistent with Polat (2020) who revealed that raters tended to be more critical and attentive when judging grammatical and lexical competences. Table 5 displays that *Obsvd Average* of structure (3.91) and *Fair Means Average* (3.94) are not significantly different. The scores of these two measures are event the same namely 4.01 for vocabulary. It is confirmed by bias logits (Table 6) where *Obs-Exp Averages* for vocabulary and structure were so mild. For vocabulary it ranges from -.02 to .03 while for structure it is .00. It indicates that rater 1 and rater 2 really paid an attention when scoring structure and vocabulary. Like Koizumi and Watanabe's finding (2021), fluency also becomes an issue to address following the present study. It has the second biggest bias logits (-.52 to .53). It indicates that the raters'

attention may have been dragged by the students' performance in such a way that they tended to overrate or underrate and shifted away from what was supposed to measure.

CONCLUSION

This study investigated rater behaviour in Kaiwa (会話) assessment with respect to three facets i.e., student, rater, and test component. Based on the analyses, it can be concluded that:

First, no significantly different scores across raters were found as confirmed by descriptive statistics analysis. Additionally, the results obtained from MFRM analyses revealed that there were different patterns of behaviour between rater 1 and rater 2. Rater 1 was more lenient than rater 2 except for pronunciation. Respective to their severity or leniency, rater 2 demonstrated more consistency in scoring pattern. In other words, she was more loyal to her stringency tendency. Rater 1 who has more experiences in Kaiwa assessment exhibited inconsistent pattern of grading, swinging from leniency to severity when it came to testing different students. With less experiences and trainings, rater 2 turned out to be more impartial and dependable. This is just in line with Koizumi and Watanabe (2021) who stated that raters are varied in the level of stringency and consistency. Some raters are more stringent while others are more generous. In the present study, rater 1 and rater 2 seemed to lean on their own interpretation

Second, both raters made biases in grading the students' performance with 15 and 14 conducts respectively. These biases occurred to 16 out of 75 students (21.3%) taking part in this study. Even more, rater 1 and rater 2 imposed biased judgements to the same students several times. Rater 1 in particular and rater 2 demonstrated unpredictable biases that befell to any students irrespective of their ability.

Third, pronunciation (-.86 to .87) logits and fluency (-.52 to .53) logits were two components that contributed the most to biases. Structure was the most objective component being scored (.00 to .01). This finding is congruent with previous studies like Polat (2020) who argued that raters tend to be attentive when judging structure. In addition, fluency could be tricky in the sense that raters could be deceived by the ratee' performance as stated by Koizumi and Watanabe (2021).

This study was certainly subject to several limitations. First, only two raters were involved. Second, both raters are non-native speakers of Japanese language. Third, it was done to freshmen whose proficiency is inadequate. Hence, for future research it is highly recommended to involve more raters and if possible, Japanese native speakers. Also, similar studies should be done with the participation of senior students whose Japanese proficiency is more advanced.

The implications of the study concerns Kaiwa assessment practices. To begin with, rater calibration should be done on a regular basis since one's ability may decline with ages. To have a similar perception on performance judgment and grading, a scoring moderation could be held. MFRM could be considered as a powerful means that helps reduce the effect of raters' bias. However, it only serves as a statistical control not to replace raters training, calibration, scoring moderation as well as rubric adjustment. Rasch measurement will provide a useful tool to model the dimensionality of Kaiwa (Japanese speaking) assessment and ascertain the reliability of its measurements. It should be kept in mind that a good assessment is a well-planned procedure that begins with the assessment construction, the to the rating process before the results are eventually used in determining candidates' performance. Each of these stages is crucial, mostly the rating process and that MFRM allows for a sounder psychometric quality.

Declaration of Conflicting Interest

The authors hereby state that there are not any conflicts of interest in this study and they will be responsible for any damage or loss inflicted.

REFERENCES

Aryadoust, V., Tan, H. A. H., & Ng, L. Y. (2019). A scientometric review of Rasch measurement: the rise and progress of a specialty. *Frontiers in Psychology*, *10*, 1-16. https://doi.org/10.3389/fpsyg.2019.02197

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: recommendations and guidelines for research. *Language Testing*, *38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences*. 4th edition. Routledge.

Engelhard, G., Wang, J., & Wind, S. A. (2018). A tale of two models: psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling*, *60*(1), 33-52.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework. *Language Sciences*, *11*(330), 1-14. https://doi.org/10.3389/fpsyg.2020.00330

JF Standard for Japanese Language Education. (2010). https://jfstandard.jp/summaryen/ja/render.do

Koizumi, R., & Watanabe, A. (2021) Rater reliability in classroom speaking assessment in a Japanese senior high school, *ARELE: Annual Review of English Language Education in Japan, 32*, 129-144. https://doi.org/10.20581/arele.32.0_129

Lane, S. (2019). Modeling rater responses processes in evaluating score meaning. *Journal of Education Measurement*, *56*(3), 653-663. https://doi.org/10.1111/jedm.12229

Lee, S. (2018). Ryūgakusei to nihonjin chūtā no gakushū katsudō: Kaiwa ni okeru kyōka ni chakumoku shite. *Gengo Bunka Kyōiku Kenkyū*, *16*, 157–176.

Linacre, J. M. (2018a). Facets Rasch measurement computer programs (Version 3.8.1). [Computer software]. Winstep.com

Linacre, J. M. (2018b). A user's guide to FACETS: Rasch-model computer programs. Winstep.com. http://www.winstep.com/facets.htm

Linacre, J. M. (2019). *Winsteps Rasch measurement computer programs User's Guide.* Winstep.com

Linacre, J. M. (2022). *A user's guide to FACETS: Rasch-model computer programs: 3.84.0 [Computer software]*. Winstep.com.

McNamara, T., Knoch, U., & Fan, J. (2019). Fairness, justice, and language assessment: *The role of measurement*. Oxford University Press

Namaziandost, E. (2019). The assessment of oral proficiency through holistic and analytic techniques of scoring: a comparative study. Frontiers Psychology. *3*, 70-82. https://doi.org/10.14744/alrj.2019.83792

Noh, M. F. M., & Matore, M. E. E. M. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.941084

Peabody, M. R., & Wind, S. A. (2019). Exploring the influence of judge proficiency on standard-setting judgments. *Journal of Education Measures, 56*.101-120. https://doi.org/10.1111/jedm.12202

Polat, M. (2020). A Rasch analysis of rater behavior in speaking assessment. *International Online Journal of Education and Teaching (IOJET)*, *7*(3). 1126-1141.

Seedhouse, P., & Nakatsuhara, F. (2018). *The discourse of the IELTS speaking test: Interactional design and practice*. Cambridge University Press.

Sultana, N. (2019). Language assessment literacy: An uncharted area for the English language teachers in Bangladesh. *Language Testing in Asia*, *9*(1), 1–14. https://doi.org/10.1186/s40468-019-0077-8

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. https://doi.org/10.1017/S0272263119000421

Weilie, L. (2018) To what extent do non-teacher raters differ from teacher raters on assessing story-telling. *Language Test Assessment*, *1*, 1-13. https://doi.org/10.23977/langta.2018.11001

Wind, S. A., & Jones, E. (2019). Not just generalizability: A case for multifaceted latent trait models in teacher observation systems. *Educational Researcher*, *48*(8), 521-533. https://doi.org/10.1016/j.asw.2022.100623

.