JLTA Journal, vol. 23: pp. 57–77, 2020

DOI: 10.20622/jltajournal.23.0_57
Print ISSN 2189-5341 Online ISSN 2189-9746

Comparison Between Holistic and Analytic Rubrics of a Paired Oral Test

Rie KOIZUMI
Juntendo University
Yo IN'NAMI
Chuo University
Makoto FUKAZAWA
University of the Ryukyus

Abstract

The current study aimed to reveal similarities and differences between a holistic and an analytic rubric used in assessing speaking performance in a paired oral test. To this end, speaking performances of 110 Japanese university students produced in paired oral interaction were evaluated by raters, holistically and analytically. The comparisons made between the two rubrics using many-facet Rasch measurement showed that both worked effectively, with the analytic rubric working slightly better in terms of a better global fit, a better test-taker and task separation, higher test-taker and task reliability, smaller standard errors, and a smaller percentage of test takers with overfit. Correlation and regression analysis indicated a strong relationship between the two (r = .84) and the Interactive communication and Fluency analytic criteria substantially explained holistic scores (adjusted R^2 = .71). Results suggest that teachers can obtain similar results with either rubric type and, if they select an analytic one, a priority would be to include Interactive communication and Fluency criteria.

Keywords: speaking assessment, rubric type, many-facet Rasch measurement

The significance of enhancing and assessing interactional competence has recently garnered special attention in the language assessment community (Galaczi & Taylor, 2018). Considering learners' future use of their target language, spoken interaction is increasingly highlighted in the English-as-a-foreign-language context in Japan, as suggested by its explicit inclusion in the Course of Study (Japanese national curriculum of English for primary and secondary schools), which has been implemented since 2020. However, English teachers in Japan generally have limited knowledge and experience in assessing spoken interaction, particularly in a classroom setting. Thus, there is an immediate need for raising their language assessment literacy related to tasks and scoring rubrics for oral interaction assessment.

One viable format in this context is a paired oral test, where two students talk or play assigned roles based on instruction cards in the second language (L2). This method has been

used to elicit a relatively natural oral interaction and a wide range of speech functions (e.g., negotiating for meaning, persuading) between two people with similar status, and it is believed to generate positive washback on students' learning (Galaczi & Taylor, 2018).

In Japan, paired oral tests have been occasionally used in practice, and research on such tests has been emerging. Nitta and Nakatsuhara (2014) examined Japanese university students' interactions in a paired oral test with and without pre-task planning time. They used an analytic rubric consisting of Fluency, Accuracy, and Complexity criteria. Negishi (2015) used holistic rubrics for assessing Japanese university students' utterances through picture description, paired, and group oral tests, comparing differences in difficulty across the three formats. Kashio et al. (2019) developed an analytic rubric, including Appropriate interaction and Appropriate response criteria, to assess oral interaction of Japanese junior and senior high school students. Matsumura and Moriya (2019) constructed an analytic rubric with Interaction, Content, Fluency, and Accuracy criteria, to assess paired interaction of Japanese university students. They examined the effectiveness of rubrics in a paired test format. Koizumi et al. (2016) developed paired oral tasks for Japanese university students and a holistic rubric. All of the above five studies suggest the feasibility of paired oral tests and the usefulness of a holistic or analytic rubric; however, no studies have developed or compared both rubric types so far. It is expected that having two rubric types will enhance the usability and expand the use of paired oral tests.

While having two rubric types will increase the number of choices for language teachers, some may wonder how they should select or use them. According to Khabbazbashi and Galaczi (2020), rigorous analyses by comparing two rubric types are limited in the L2 speaking field, which contrasts sharply with the L2 writing field, where many such studies have been conducted (see, for example, Barkaoui, 2011; Ono et al., 2019). Empirical research on rubric comparison for assessing L2 speaking ability, especially the oral interactive ability, would provide information on which works better and how they are related, and this would lead to useful suggestions on the rubric selection and use. The current study compares holistic and analytic rubrics with this need in mind.

Literature Review

Relationships Between Holistic and Analytic Rubrics

According to Brookhart (2013), "a rubric is a coherent set of criteria for students' work that includes descriptions of levels of performance quality on the criteria" (p. 4). A rubric can be used for evaluating learners' performances and giving detailed feedback to learners and teachers. When raters evaluate performance, they consider a range of criteria jointly and produce one global score in a holistic rubric, or they produce multiple scores for different criteria separately by using an analytic rubric. Typically, both rubrics use a grid (form) with descriptors, or descriptions of language performance or ability, for each level inside the grid. The development and validation of rubrics are essential due to the fact that a rubric represents an operational definition of a test construct and plays a central role in scoring (McNamara,

1996). A rubric is also termed as a rating scale; a criterion is also called a category.

Holistic and analytic rubrics have their own characteristics, which are typically opposite to each other. Previous research (Brown, 2012) suggests that, although a holistic rubric is easier and more efficient to use, it lacks the diagnostic information to help improve future learning and teaching that an analytic rubric offers. However, the latter takes more time to score.

Table 1 summarizes previous studies that examined correlations between holistic and analytic rubric scores. The results suggest that correlations are strong across speaking task types, such as speech, interviews, monologues, and across different types of learners, indicating that the two scoring methods produce similar results. For example, Zhang (2019) compared three types of rubrics, two of which are relevant to the current study. She used 166 speech data from the College English Test-Spoken English Test Band 4, which consisted of three task types: reading a text aloud, individual presentation, and pair discussion. She asked six raters to evaluate the speech samples using (a) holistic rubrics developed for each task type (termed "task-based holistic rubric") and (b) five analytic criteria for scoring all the tasks (termed "test-based analytic rubric"). The correlation between the holistic and analytic total scores was very high, at .92.

While correlations were consistently strong in previous studies, a detailed analysis of each scale has revealed subtle differences, especially by analyzing measurement properties using many-facet Rasch measurement (MFRM). Zhang (2019) showed that the analytic rubric discriminated between test-takers' different proficiency levels better than the holistic rubric, as evidenced by the higher test-taker separation. A better distinction between test takers in an analytic rubric was also found in L2 writing studies that used one prompt (Barkaoui, 2011, with 168 learners and 60 raters; Wiseman, 2012, with 60 learners and 5 raters). By contrast, Khabbazbashi and Galaczi (2020) reported better results in the holistic rubric when they asked 10 raters to evaluate 200 test-takers' monologues: a clearer distinction between test takers, higher test-taker reliability, and a better fit to the Rasch model. Opposing results in two L2 speaking studies suggest that there may be cases where holistic rubrics work better. As such, further studies would provide more insights into holistic vs. analytic differences. In particular, a specific focus on a paired oral task type with several tasks would uncover more details that would be useful in selecting holistic and analytic rubrics. The current study focuses on this aspect (see Table 1). Currently, Zhang (2019) seems to be the only study that included a paired oral task, but this was one of the three task types examined there.

Analytic Criteria Predicting Holistic Scores

Another aspect to be examined in terms of relationships between holistic and analytic rubrics is what analytic criteria substantially contribute to holistic scores. This question can be rephrased as follows: What is assessed in holistic scores? What is the construct of holistic scores? McNamara (1990) used multiple regression analyses to examine relationships between holistic scores and five analytic criterion scores in the speaking section in the Occupational English Test (OET) among 192 and 198 test takers. The data were analyzed separately across

Table 1

Previous Studies That Examined Correlations Between Holistic and Analytic Rubric Scores

Study	Test takers	Raters	Task	Analytic rubric	*r* [95% CI]
Xi & Mollaun (2006)	140 university students	14 raters	Semi-direct talk	3 criteria	.92 [.89, .94]
Khabbazbashi and Galaczi (2020)	200 test takers	10 raters	Semi-direct talk	4 criteria	.80 [.76, .84]
Naito (1995)	11 Japanese high school students	10 English teachers[a]	Public speech	3 criteria	.89 [.62, .97]
Fukuda (2018)	10 Japanese junior high school students	52 English teachers[a]	Public speech	5 criteria	.83 [.42, .96]
Aso (2000)	36 Japanese high school students	10 English teachers[a]	Examiner-led interview	6 criteria	.66 [.42, .81]
Kobayashi (2005)	9 Japanese high school students	14 English teachers[a]	Examiner-led interview	8 criteria	.57 [−.15, .90] to .96
Chuang (2009)	5 Taiwanese college students	62 Taiwanese college teachers	Examiner-led interview	5 criteria	.97[b] [.61, .99]
Metruk (2018)	50 Slovak university students	2 Slovak university teachers	Examiner-led interview	4 criteria	.86[b] [.77, .92]
Zhang (2019)	166 CET-SET4 test takers	6 college English teachers	3 tasks: e.g., read aloud, paired discussion	5 criteria	.92 [.89, .94]
Negishi (2011)	135 Japanese learners of English[c]	10 Japanese teachers of English	Group oral	5 criteria	.99[b] [.99, .99]
Current study	110 Japanese university students	3 to 4 English teachers	Dialogue (paired)	4 criteria	--

Note. CI = confidence interval, which was calculated using the online website (http://vassarstats.net/rho.html). CET-SET4 = College English Test-Spoken English Test Band 4. [a]Including native speakers of English and Japanese teachers of English. [b]Value calculated using the raw data in the article. [c]Junior and senior high school students and university students.

two years. He found that holistic scores (Overall communicative effectiveness) were substantially explained by Resources of grammar and expression (Grammar; R^2 = .678 and .695, respectively), followed by Fluency (additional R^2 = .086 to .062). The large contribution of a grammatical criterion to holistic scores was not expected, as the OET had a communicative focus, with grammar less highlighted unless it hampered communication. Using the same speaking test, Iwashita and Grove (2003) conducted the same analysis with the data from 7,347 test takers. They found that the best predictor of holistic scores was Fluency (β = .29, R^2 not reported), followed by Grammar (β = .24), although it is difficult to compare their study with McNamara (1990) since they did not report R^2. They wrote that the larger contribution of Fluency rather than Grammar to holistic scores may have arisen due to "different characteristics in the populations of test-takers" and changes in rater orientations, in which "grammatical and lexical accuracy is no longer treated as the most important aspect of speaking performance" (p. 31).

The speaking test in McNamara (1990) and Iwashita and Grove (2003) used an interview format in which an interlocutor elicits a test-taker's talk, using role-play tasks. Although their analytic criteria did not include an interaction-related one, holistic scores may have reflected the quality of interaction.

Paired oral tasks and group oral tasks have been found to elicit a variety of interactive language functions and to assess test-takers' interactive ability or interactive competence more than interlocutor-led interviews (Galaczi & Taylor, 2018). Thus, it is possible to predict that interaction would be the main test construct in a paired oral, that holistic scores would reflect interaction much better, and that interaction-related criterion scores would correlate with holistic scores more than other criterion scores. So far, Negishi (2011) appears to be the only study that provides data to investigate this prediction. She asked 10 Japanese teachers to evaluate a group oral discussion of 135 Japanese learners of English using a holistic rubric and an analytic rubric with five criteria (i.e., Range, Accuracy, Fluency, Interaction, and Coherence). To examine if Interaction scores have a stronger correlation with holistic scores than other criterion scores, we reanalyzed the data in her dissertation, averaging 10 raters' ratings. We found very strong correlations of holistic scores with all analytic scores (i.e., analytic total scores and each of the analytic criterion scores; r = .98 to .99). However, these strong correlations may have been derived due to the fact that raters evaluated the group oral consecutively holistically and analytically, in this order, and holistic scores that they had just assigned might have led to strong correlations (i.e., halo effects). Although McNamara (1990) used the same rating procedure, scores in each analytic criterion showed relatively different patterns regarding correlations with holistic scores. This may be because McNamara's (1990) raters listened to 15-minute talk interviews and two role plays and had two opportunities to reconsider their scores and to give separate scores. On the other hand, Negishi's (2011) raters listened to 5-minute group interactions and decided on the holistic score first, followed by the analytic ones, and they may not have been able to give scores independently. Thus, the different results might have been due to the differences in detailed procedures. The current study will

ensure separate scoring across rubric types and examine relationships between holistic and analytic scores of a paired oral test.

Purpose of the Study

We aim to compare a holistic rubric with an analytic rubric of a paired oral test to examine their measurement properties and relationships between them. We specifically address the following research questions, using MFRM, correlation, and regression.

Research question 1: How different are holistic and analytic rubrics in terms of measurement properties?

Research question 2: How are holistic and analytic rubric scores correlated?

Research question 3: What analytic criteria substantially contribute to holistic scores?

Method

Participants

Students at three Japanese universities (N = 110) with low to intermediate English proficiency levels took a paired oral test. Their native language was Japanese. Most (80.00%) of the students majored in science and engineering; the rest majored in medicine (16.36%) and global studies (3.64%). Most were males (71.82%). This data is partly taken from Koizumi et al. (2016), which had 190 participants. For this study, however, the number of participants was reduced to ensure double ratings for all the participants and enough participants for each task.

Paired Oral Test

The test was designed for classroom use with the Japanese context in mind. Students paired themselves off in class and talked for two to three minutes, following the instructions on the task card (see Koizumi et al., 2016). Initially, the test consisted of a warm-up task, seven guided role-play tasks (e.g., exchanging information, invitation and polite refusal, making suggestions), and four unguided discussion tasks (e.g., talking about hobbies, deciding what to take on a trip). However, only six role-play and four discussion tasks were used for analysis. There was no planning time before they started talking. As part of the instruction in an English class, students completed 3 to 10 tasks. The number of tasks varied, since the time spent on a pair test in a class was different. A pair's interaction was recorded separately for each task using a voice recorder.

Rubrics

We used two assessment rubrics: a holistic rubric and an analytic rubric. The holistic rubric was developed by restructuring elements from Nakatsuhara's (2013) analytic rubric. It requires raters to primarily consider task fulfillment, interactive communication, and fluency, on a scale of one to three (see Appendix A and Koizumi et al., 2016 for details). The analytic rubric had four criteria with three levels: Pronunciation & intonation (Pronunciation, hereafter), Grammar & Vocabulary, Fluency, and Interactive communication (see Appendix B). It was

developed based on Nakatsuhara's (2013) 6-level analytic rubric with five criteria. We combined the "Grammar" and "Vocabulary" criteria and decreased the number of levels from six to three. This served not only to enhance practicality but also to reflect authentic spoken language, based on Römer's (2017) statement of "the inseparability of lexis and grammar" (p. 477) and suggestion to combine grammar and vocabulary as one criterion.

Scoring Procedures

Each pair's recorded interaction was scored by two raters. In the holistic scoring, four raters were involved. In the analytic scoring, three raters out of four were involved. Each paired talk was marked twice, first holistically and next analytically, at separate times. There was an interval of at least two years between the holistic and analytic scoring (in 2016 and 2018), and raters did not recall the scores they previously assigned. Therefore, we considered the two types of scoring to be unlikely to be affected by each other.

The four raters involved in this study had L2 English teaching experience of more than 10 years. A rater training was first conducted. Raters familiarized themselves with the holistic rubric and evaluated 10 participants' recordings. They discussed the discrepancies in their ratings for five to eight hours. Afterwards, they independently scored each student, listening to the assigned recordings. Later, the same procedure was taken with the analytic rubric.

Analysis

We analyzed the assigned ratings using MFRM with Facets (Ver. 3.83.1; Linacre, 2020a; see McNamara et al., 2019, for details). We included three facets (i.e., test-takers, tasks, and raters) and four facets (i.e., test-takers, tasks, raters, criteria) for the holistic and analytic data, respectively. The rating scale model was used for the holistic ratings, whereas the partial credit model was used for the analytic ratings, with the criteria facet modeled as partial credit because each criterion was assumed to function independently.

The fit to the model of each facet was evaluated using Infit mean squares between 0.50 and 1.50 (Linacre, 2020b). Values of less than 0.50 (overfit) mean that response patterns are too predictable; however, an element (e.g., task) is "not degrading" measurement. Values of more than 1.5 to 2.0 (underfit) show that patterns are unpredictable but an element is "not degrading" measurement. Values of more than 2.0 (underfit) are also unpredictable, with an element "distorting or degrading the measurement system" (p. 285). The appropriateness of the rubrics was judged based on Bond and Fox (2015).

The initial analysis of the analytic ratings showed that the Pronunciation criterion required modification. Level 1 was used only once (0.0005%) since most of the participants' talk was comprehensible. We combined Levels 1 and 2, resulting in Pronunciation having two levels. After this change, we had 2,109 holistic data points (i.e., valid responses used for estimation) and 8,107 analytic data points; results for these data are reported below.

To answer Research question 1, we examined a global model fit of the data, the fit of each facet to the Rasch model, as well as other aspects. We included students who obtained full

marks (holistic: n = 7, 6.36%; analytic: n = 3, 2.73%) and had no students with zero marks. To answer Research questions 2 and 3, we employed Pearson product-moment correlations and multiple regressions using JASP (Ver. 0.12.1; https://jasp-stats.org/). For this purpose, we used test-taker ability estimates from MFRM, since the number of tasks varied across test takers, which made it difficult for us to use the raw scores. We conducted MFRM six times, once with the holistic rubric, once with four analytic criteria, and four times with each criterion separately.

We conducted hierarchical linear multiple regression with holistic estimates as a dependent variable and four analytic criterion estimates as independent variables. We first checked the assumptions of multiple regression (Takaki, 2017) and confirmed that all the five required assumptions were satisfied. First, the number of participants (N = 110) was larger than the number required for multiple regression: 50 + 8 × the number of independent variables (82 = 50 + 8 × 4) and 104 + the number of independent variables (108 = 104 + 4). Second, multicollinearity was not detected, as the tolerance values ranged from .15 to .47; multicollinearity is suspected in cases of tolerance values of .10 or lower. Third, outliers were not detected; Cook's distance values (which should be less than 1.00) were 0.11 or lower. Fourth, independence of residuals was confirmed; the Durbin-Watson statistic (which should be around 2) was 1.84. Last, normality of residuals, homoscedasticity of residuals, and linearity of residuals were confirmed with graphs.

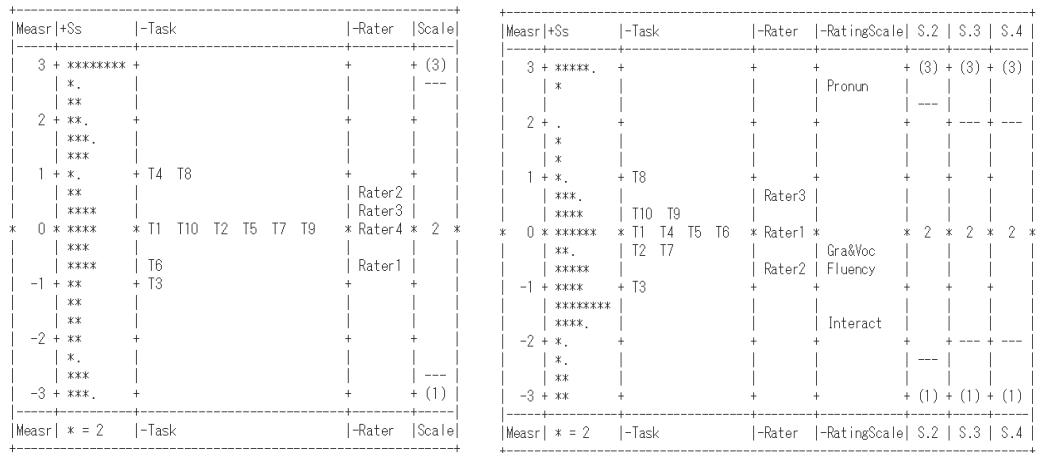

Figure 1. Wright maps in the holistic rubric (left) and analytic rubric (right). S.2 = Grammar & Vocabulary (Gra&Voc); S.3 = Fluency; and S.4 = Interactive communication (Interact). S.1 (Pronunciation: Pronun) was not shown because it had only two levels. T1 = RP_club (Role-play task with a topic of clubs); T2 = RP_dinner; T3 = D_hobbies (Discussion task with a topic of hobbies); T4 = D_trip; T5 = RP_job; T6 = RP_movie; T7 = D_friends; T8 = D_date; T9 = RP_toothache; T10 = RP_driving. This applies to Table 3.

Results

Many-Facet Rasch Measurement (MFRM)

Figure 1 shows Wright maps that indicate how each facet was related to one another in the holistic and analytic rubrics. They suggest that, in both rubrics, the distributions of test-taker ability estimates and task difficulty estimates generally fit with each other, and the current data is appropriate for our research purpose (in the holistic scoring, test taker: $M = 0.53$, $SD = 2.53$; task: $M = 0.00$, $SD = 0.60$; in the analytic scoring, test taker: $M = -0.07$, $SD = 2.20$, task: $M = 0.00$, $SD = 0.43$). Table 2 shows a summary of the overall measurement properties of holistic and analytic rubrics.

As seen in Table 2, a global model fit was examined using standardized residuals obtained from the Unexpected responses in the output table (Linacre, 2020b, p. 176). We found that results in both rubrics showed a satisfactory global model fit. However, it was slightly better in the analytic rubric in terms of standardized residuals of beyond ±2 (4.74% vs. 1.23% in the order of holistic and analytic rubrics). By contrast, it was slightly worse in the analytic rubric, although similar to the holistic rubric in terms of standardized residuals of beyond ±3 (0.33% vs. 0.58%).

In terms of test-taker ability, although both rubrics had appropriate measurement properties, the analytic rubric had slightly better results, evidenced by a higher separation and strata, higher reliability, and a smaller percentage of test takers with overfit. A higher separation, strata, and reliability mean that the analytic rubric could better differentiate between test takers. In terms of underfit, both rubrics were similar.

Regarding task difficulty, both rubrics produced appropriate measurement properties. The analytic rubric had slightly better results, due to a higher separation and strata, and higher reliability, which suggests that the analytic rubric better discriminated between tasks. In terms of underfit and overfit, both rubrics were similar, with no problematic tasks.

Regarding rater severity, the interpretation of the results differs depending on how scores are analyzed. When MFRM is used, high separation and reliability are not an issue as long as raters fit the model (which they did, as described below) because MFRM can adjust rater severity differences. Thus, the results for both rubrics were appropriate. In contrast, when raw scores are used, it is better to have a lower rater separation and strata and lower rater reliability. For this reason, the analytic rubric had slightly worse outcomes, with a higher separation and strata and higher reliability. This indicates that rater severity differed more in the analytic rubric, which could lead to slightly greater variation due to rater differences. The results also show that the analytic rubric better discriminated between rater differences. In terms of underfit and overfit, both rubrics reported the same, with no raters beyond the expectation of the Rasch model. Besides, the agreement among raters was sufficiently high and also slightly higher than the one predicted by the Rasch model, in both rubrics (77.2% vs. 78.0%, in holistic and analytic rubrics, respectively; see Table 2, *Note*).

Concerning the functioning of the two rubrics, there are five requirements to consider when deciding whether a rubric has worked well (Bond & Fox, 2015): (a) Each level has more

than 10 observations; (b) level difficulty estimates (i.e., Average measures) increase as levels increase (c) fit statistics (i.e., outfit mean squares) are less than 2.0; (d) Rasch-Andrich thresholds measures increase as levels increase; (e) distances (i.e., differences between thresholds) are between 1.4 and 5.0 logits; (f) there is a clear peak for each level in the probability curve. Results were positive in both rubrics in terms of (a), (b), (d), and (f). However,

Table 2

Comparisons Between Many-Facet Rasch Analysis Results

	Criteria	Holistic	Analytic
Global model fit	Standardized residuals (SRs) outside ±2 should be less than 5%. SRs outside ±3 should be less than 1%.	outside ±2: 4.74% (100/2,109) outside ±3: 0.33% (7/2,019)	1.23% (100/8,107) 0.58% (47/8,107)
Test-taker ability			
Variation	Separation & Strata: the higher, the better	3.39, 4.85 [1.70, 2.43]	5.05, 7.07 [2.92, 4.08]
	Reliability: the higher, the better	.92	.96
Fit statistics[a]	Underfit (Infit MSs > 2.00)	1.82% (2/110)	0.91% (1/110)
	Underfit (Infit MSs > 1.50)	8.18% (9/110)	9.09% (10/110)
	Overfit (Infit MSs < 0.50)	6.36% (7/110)	2.73% (3/110)
Task difficulty			
Variation	Separation & Strata: the higher, the better	3.44, 4.92 [1.72, 2.46]	4.87, 6.82 [2.81, 3.94]
	Reliability: the higher, the better	.92	.96
Fit statistics[a]	Underfit (Infit MSs > 1.50)	0%	0%
	Overfit (Infit MSs < 0.50)	0%	0%
Rater severity[b]			
Variation	Separation & Strata: the higher, the worse[c]	4.12, 5.82 [2.06, 2.91]	9.32, 12.76 [5.38, 7.37]
	Reliability: the higher, the worse[c]	.94	.99
Fit statistics[a]	Underfit (Infit MSs > 1.50)	0%	0%
	Overfit (Infit MSs < 0.50)	0%	0%

Note. [ ] = Standardized separation & strata, which was calculated based on Zhang (2019), fusing the formula: [separation (or strata)]/[√(the number of raters)] to balance out a different numbers of raters. MSs = mean squares.

[a]A smaller percentage of underfit and overfit is better. [b]Exact agreements in the holistic rubric = 77.2% (> Expected = 65.6%). Exact agreements in the analytic rubric = 78.0% (> 69.8%). [c]The higher the value, the worse the situation when raw scores are used. When MFRM is used, high separation and reliability are not problematic as long as raters fit the Rasch model.

requirements (c) and (e) were not satisfied. First, (c) the fit statistic was more than 2.0 (i.e., 2.2) at Level 2 in the Interactive communication criterion, in the analytic rubric. This suggests that Interactive communication received some unexpected responses at Level 2. Second, (e) distances were more than 5.0 in the holistic rubric and in Grammar & Vocabulary in the analytic rubric (i.e., both 5.26). This means that the distances between Levels 1 and 3 were too wide, suggesting that increasing the number of levels would provide a better discrimination and more test information. One way of revising may be to make Level 2 more difficult and/or make Level 3 easier by modifying level descriptors or rater training. For the current test, as this paired oral test is intended to be used in classroom speaking assessments, we prefer to retain 3 levels to maintain practicality in terms of scoring so that teachers can use the rubrics more easily.

To delve into the underfit aspect of the Interactive communication analytic criterion at Level 2, we examined unexpected responses with standardized residuals of 2.00 or more and found that there were 20 unexpected responses related to Interactive communication. Out of 20, 17 (85.00%) had Level 2 observed scores, with the remaining 3 having Level 1 observed scores. The 20 responses were judged as unexpected due to the observed scores that raters gave being much lower than the ones expected, which were the scores that a certain test taker with this ability would be likely to receive, according to the Rasch model. This suggests that raters, occasionally, assigned harsher scores when evaluating Interactive communication, especially at Level 2. Therefore, Interactive communication may require revision for the Level 2 descriptors and/or more intense rater training regarding it.

Relationships Between Holistic, Analytic Total, and Analytic Criterion Scores

Table 3 shows distributions of the six types of scores (or estimates) that were derived from six runs of MFRM: holistic, analytic total, Pronunciation, Grammar & Vocabulary, Fluency, and Interactive communication scores. In terms of comparisons between holistic and analytic rubrics, analytic scores showed a better measurement property and had a smaller mean of standard errors (*M SE* = 0.34), with a smaller standard deviation of standard errors (*SD SE*

Table 3
Summary Statistics of Test-Taker Ability Estimates of Six Types of Scores

	M measure	*SD* measure	*M SE*	*SD SE*	Separation	Strata	Reliability
Holistic	0.53	2.53	0.64	0.64	3.39	4.85	.92
Analytic	−0.07	2.24	0.34	0.27	5.05	7.07	.96
Pronunciation	−2.43	3.06	1.64	0.43	1.50	2.33	.69
Gra&Voc	0.16	2.32	0.63	0.29	3.22	4.62	.91
Fluency	0.70	2.45	0.64	0.35	3.19	4.58	.91
Interact	1.61	2.12	0.65	0.36	2.69	3.92	.88

Note. *SE* = standard error. Exact agreements: Pronunciation = 92.1% (< Expected = 93.9%); Gra&Voc = 76.4% (> 67.6%); Fluency = 71.2% (> 64.0%); Interact = 72.3% (> 65.0%).

= 0.27) than holistic scores (*M SE* = 0.64 and 0.64, respectively). All the scores in the holistic and analytic rubrics had high reliability except for Pronunciation scores, which had relatively low reliability (.69) and whose standard errors were larger than other scores (*M SE* = 1.64). This seems to be related to the fact that Pronunciation has only two levels and has less information than other analytic criteria. Figure 2 shows that Pronunciation had a different distribution from other scores.

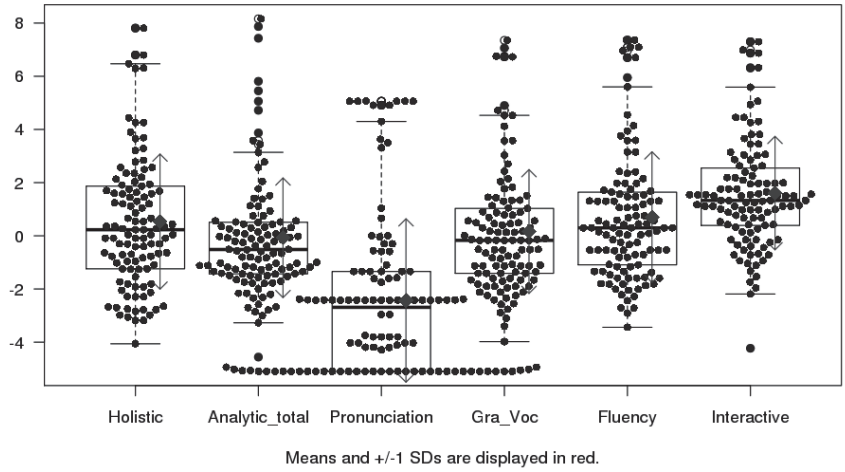

Figure 2. Box-and-whisker and beeswarm plots that show distributions of the six variables. The figure was derived using the website (http://langtest.jp/). Gra_Voc = Grammar & Vocabulary. Interactive = Interactive communication. The same applies to Figure 3 and Table 4.

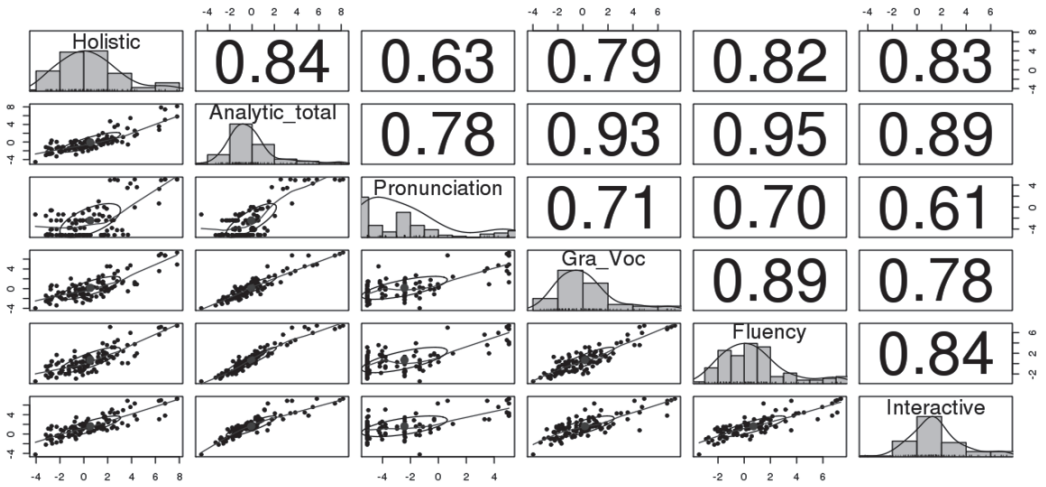

Figure 3. Histograms, scatterplots, and correlations between holistic scores, analytic total scores, and analytic criterion scores.

As can be seen in Figure 3, the correlation between holistic and analytic scores was strong, at r = .84 [95% confidence interval: .78, .89]. This suggests that these scores were strongly related but not identical. To examine the differences across the rubrics, we conducted simple regression and identified six test takers with large differences between the scores (i.e., those with absolute standardized residuals of more than 2.0). A closer examination showed that relatively large gaps arose due to the fact that each rubric considers overlapping but different aspects: The holistic rubric considers task fulfillment together with fluency and interactive communication, whereas the analytic rubric fails to consider task fulfillment. We found two students who had higher analytic than holistic scores, as they did not try hard to achieve task requirements; however, they gained high analytic scores, especially in Interactive communication, due to their inherent high ability. Another two students had very high fluency but did not accomplish the tasks well since their talk was irrelevant to the assigned topic, which resulted in obtaining lower holistic scores. Another two students had higher holistic scores since they both achieved the tasks well, but the language quality was not superb, for example, in Pronunciation and Interactive communication.

Next, we examined what analytic criteria explain the holistic scores. Applying four regression analyses showed the following results, as summarized in Table 4. First, when four analytic criterion scores were entered as independent variables into the regression equation, they explained approximately 75% (adjusted R^2 = 74.9%) of holistic scores. However,

Table 4

Regression Results for the Holistic Scores as an Dependent Variable

Cri	Variable	*B*	95% CI for *B*	*SE B*	β	*t*	*p*	R^2
1	(Intercept)	−1.06	[−1.40, 0.73]	0.17	--	−6.31	< .001	.696 (.693)
	Interactive	0.99	[0.87, 1.12]	0.06	.83	15.72	< .001	
2	(Intercept)	−0.70	[−1.04, −0.35]	0.17	--	−4.02	< .001	.748 (.743)
	Interactive	0.58	[0.37, 0.79]	0.11	.48	5.42	< .001	ΔR^2 = .052
	Fluency	0.43	[0.25, 0.62]	0.09	.42	4.69	< .001	
3	(Intercept)	-0.56	[−0.92, 0.20]	0.18	--	-3.07	.003	.758 (.751)
	Interactive	0.55	[0.34, 0.75]	0.11	.46	5.15	< .001	ΔR^2 = .010
	Fluency	0.26	[0.01, 0.51]	0.12	.25	2.09	.039	
	Gra_Voc	0.24	[0.01, 0.46]	0.11	.22	2.07	.041	
4	(Intercept)	−0.48	[−0.97, 0.01]	0.25	--	−1.93	.056	.758 (.749)
	Pronunciation	0.03	[−0.09, 0.14]	0.06	.04	0.49	.623	ΔR^2 =.000
	Gra_Voc	0.22	[−0.01, 0.45]	0.12	.20	1.86	.066	
	Fluency	0.25	[−0.004, 0.50]	0.13	.24	1.95	.054	
	Interactive	0.54	[0.33, 0.75]	0.11	.46	5.12	< .001	

Note. Cri = No. of criteria entered. ( ) = Adjusted R^2. ΔR^2 = change in R^2.

Interactive communication was the only statistically significant predictor (β = .46). Second, when analytic criterion scores were gradually entered into the regression analysis, Interactive communication predicted holistic scores at about 69% (adjusted R^2 = 69.3%; β = .83), Fluency predicted them at an additional 5.2%, and Grammar & Vocabulary predicted them at an additional 1.0%. Therefore, the contribution of Interactive communication was the largest, followed by Fluency and marginally by Grammar & Vocabulary. Because the core construct of a paired oral speaking test is Interactive communication, this result can be interpreted as positive evidence for the validity of interpretations based on holistic scores.

Discussion

Regarding Research question 1 (How different are holistic and analytic rubrics in terms of measurement properties?), the overall results using MFRM suggest that both rubrics function effectively. However, the analytic rubric worked slightly better in four aspects. First, global fit was slightly better in the analytic rubric than in the holistic rubric, as the percentage of unexpected responses from MFRM in the analytic rubric (1.23%) was smaller. Second, the analytic rubric had slightly higher test-taker and task reliability (.96 and .96, respectively) and had a smaller mean of standard errors (*M SE* = 0.34), with a smaller standard deviation of standard errors (*SD SE* = 0.27). Third, the analytic rubric could separate test takers more finely, into at least five test-taker groups (5.05), and differentiate test tasks better, into about five different task difficulty levels (4.87). Fourth, the percentage of test takers with overfit was smaller (2.73%). On other aspects, such as the percentage of underfit and overfit in test takers, tasks, raters, high rater agreement, and most (4 out of 6 requirements) of the rating scale functioning, the holistic and analytic rubrics worked similarly. On the remaining point, the holistic rubric worked better. For rater variation, the analytic rubric revealed a finer differentiation in rater severity, dividing it into nine different levels (9.32). As explained above, this may be an issue when raw scores are used. Furthermore, regarding rubric functioning, Interactive communication in the analytic rubric had a high underfit value (Outfit mean squares = 2.2), which indicates areas for improvement but also suggests that the analytic rubric can identify such subtle irregularities.

To focus on the results of test-taker ability estimates, while the superior results of the analytic rubric were consistent with Zhang (2019), Barkaoui (2011), and Wiseman (2012), they were different from Khabbazbashi and Galaczi (2020), in which the holistic rubric showed a higher test-taker separation, higher test-taker reliability, as well as a better fit to the Rasch model than the analytic rubric.

To examine this difference, a further classification may seem helpful, as suggested by Khabbazbashi and Galaczi (2020). When there were multiple tasks (e.g., three tasks) in a test, one scoring method is to assess responses in all the tasks holistically and give one holistic score for the whole test. Imagine a situation where raters give a holistic score after listening to Tasks 1 to 3. The rubric used for this method can be termed as a *whole* holistic rubric. By contrast, raters may assess responses in each task holistically and give a holistic score for each task (e.g.,

raters give a holistic score after listening to Task 1 and repeat the same actions with Tasks 2 and 3, resulting in three holistic scores for each test taker). This rubric can be named a *part* holistic rubric (a *part* means a task, in this hypothesical case). Furthermore, the analytic rubric can also be divided into (a) *whole* analytic and (b) *part* analytic. (a) A *whole* analytic rubric considers all task performances together and gives analytic scores to the whole test. Suppose an analytic rubric comprises four analytic criteria; raters give four analytic criterion scores after listening to Tasks 1 to 3. By contrast, (b) a *part* analytic rubric considers each task performance, giving analytic criterion scores to each task (e.g., raters give four analytic criterion scores to Task 1, and iterate the same procedure for Tasks 2 and 3, resulting in 12 scores [4 criteria x 3 tasks]). Table 5 shows how rubrics in previous studies can be categorized into part or whole. Overall, this suggests that an analytic rubric basically works better; however, a part holistic rubric could be a better choice than a whole analytic rubric in some contexts.

Table 5
Summary of Previous and Current Results

Studies	L2 Skill	Holistic	Analytic	Overall results of test-taker ability
Current study	Speaking	Part	Part	The part analytic rubric was better.
Khabbazbashi and Galaczi (2020)	Speaking	Part	Whole	The part holistic rubric was better.
Zhang (2019)	Speaking	Part	Whole	The whole analytic rubric was better.
Barkaoui (2011)	Writing	NA	NA	The analytic rubric was better.
Wiseman (2012)	Writing	NA	NA	The analytic rubric was better.

Note. NA = 1 prompt was used and part vs. whole distinction is not relevant.

To explore what type of contexts are related, two studies using a part holistic rubric and a whole analytic rubric (Khabbazbashi & Galaczi, 2020; Zhang, 2019) were compared. One large difference between the two studies is that Zhang (2019) included three very different tasks (i.e., reading a text aloud, individual presentation, and pair discussion), whereas Khabbazbashi and Galaczi (2020) included four relatively similar monologue tasks (i.e., answering questions on personal topics, describing and comparing two pictures, answering questions on a familiar topic, and a long talk on an abstract topic). The scoring of similar tasks (i.e., parts) may have enabled raters to detect minor differences and differentiate test-takers' ability slightly better in Khabbazbashi and Galaczi (2020). By contrast, responses in the different tasks usually show different speech characteristics, and raters may have difficulty in gaining a consistent analytic view as a whole. The types of tasks included in a test may change which works better: a part holistic rubric or a whole analytic rubric.

Concerning Research question 2 (How are holistic and analytic rubric scores correlated?), there was a strong correlation between the two rubric scores (r = .84). Further analysis suggested that minor differences between the two rubric scores arose due to a different aspect included in the holistic rubric (i.e., task fulfillment). The strong correlations between the

holistic and analytic scores accord well with the previous studies summarized in Table 1.

For Research question 3, we examined the analytic criteria that substantially contribute to holistic scores. Multiple regression analysis showed that while three analytic criteria (i.e., Interactive communication, Fluency, and Grammar & Vocabulary) explain about 75% of holistic scores, Interactive communication predicts about 69%, with an additional 5% explained by Fluency. The result that Interactive communication was the largest predictor of holistic scores shows that the test construct of holistic scores in a paired oral test primarily consists of interactive communication. This argument is in line with Galaczi and Taylor (2018), who considered interaction as a central test construct in paired oral tests.

The finding that 74% of the holistic scores were explained by Interactive communication and Fluency suggests that these two criteria would be strong candidates when selecting an analytic rubric for paired oral tests. Some may worry about the reduction of reliability from the current four to two analytic criteria; however, in fact, the test-taker reliability decreased very little (.93 for two criteria vs. .96 for the four), which does not seem to be much of a concern.

The maximum percentage that the analytic criteria were able to predict was 75%, by using Interactive communication, Fluency, and Grammar & Vocabulary. The remaining 25% may include a task fulfillment aspect along with measurement errors, as divergences between holistic and analytic scores could be attributed to the lack of task fulfillment in an analytic rubric. The phenomenon of task fulfillment (task achievement) criterion lacking in analytic scores, causing divergence in scores across holistic and analytic rubrics, was also observed in Khabbazbashi and Galaczi (2020). A future rubric revision may need to consider including task fulfillment as one of the analytic criteria.

Conclusion

The current study compared scores derived from using a holistic rubric with those in an analytic rubric and found that both rubrics worked effectively, with the analytic rubric working slightly better in terms of a better global fit, a better test-taker and task separation, higher test-taker and task reliability, smaller standard errors, and a smaller percentage of test takers with overfit. Additionally, the two rubric scores were strongly correlated, and Interactive communication in an analytic rubric substantially explained holistic scores.

While the current results help extend insights in this field, there are two limitations that need to be kept in mind. First, the current study used four raters in the holistic assessment and three raters in the analytic one. Although there was a long-time interval (at least two years) between the holistic and analytic rating sessions, one additional rater might have affected the results. Second, the context of the current study, such as the type of participants, the paired oral test format, raters, rubrics, and their descriptors, may have affected the outcome. For example, raters were all L1 Japanese teachers of English. More research in different contexts should be conducted in the future.

The practical implications derived from this study are as follows. Teachers often wonder which rubric should be used in their speaking assessment, and in the case of using an analytic

rubric, what criteria should be used. The current study suggests that holistic and analytic rubrics produce similar results overall, and that if teachers prefer to use a simple analytic rubric in assessing speaking performances in paired oral tests, they can select interactive communication and fluency (possibly with task fulfillment) as analytic criteria. They can maintain high reliability with adequate separation of test takers, while being able to provide diagnostic feedback to students, as compared with a holistic rubric.

However, there is one caveat to be considered. We used 10 tasks in a test, but it is rare for a paired oral test to include such a large number of tasks at one time. Thus, the conclusion that an analytic rubric with two criteria can explain much of the holistic score (about 75%) while maintaining high reliability may be based on the 10-task results, and teachers may need to use more criteria when they use a fewer number of tasks. To examine the effects of the number of tasks on scores, future studies should include generalizability theory in their analyses (Brennan, 2001) to further specify the number of tasks, raters, and criteria required to maintain certain degrees of reliability. Additional qualitative analyses on rater performance and perception and other aspects would also reveal similarities and differences across rubric types.

Acknowledgments

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 and 20K00894. We would especially like to thank Yumi Koyamada and two reviewers for their great assistance.

References

Aso, Y. (2000). A comparison of holistic and analytic scorings for oral interview tests. *Annual Review of English Language Education in Japan, 11,* 131–139. https://doi.org/10.20581/arele.11.0_131

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279–293. https://doi.org/10.1080/0969594X.2010.526585

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

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Association for Supervision and Curriculum Development.

Brown, J. D. (Ed.). (2012). *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages*. National Foreign Language Resource Center, University of Hawai‘i at Mānoa.

Chuang, Y.-Y., (2009). Foreign language speaking assessment: Taiwanese college English teachers’ scoring performance in the holistic and analytic rating methods. *The Asian EFL Journal*, *11*(1), 152–175. https://www.asian-efl-journal.com/main-journals/foreign-language-speaking-assessment-chinese-taiwanese-college-english-teachers-scoring-

performance-in-the-holistic-and-analytic-rating-methods/

Fukuda, N. (2018). Eigo supichi kontesuto notameno bunsekiteki hyoka nikawaru zentaiteki hyoka no jitsuyo no kanosei [The possibility of implementing a holistic evaluation instead of an analytic evaluation for the English speech contest]. *Research Reports of National Institute of Technology, Nagaoka College*, *54*, 18–24. https://doi.org/10.24806/rrnitnc.54.0_18

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Iwashita, N., & Grove, E. (2003). A comparison of analytic and holistic scales in the context of a specific-purpose speaking test. *Prospect*, *18*(3), 25–35.

Kashio, F., Ueda, T., Minami, T., & Takemoto, K. (2019). Jiki gakushushidoyoryo niokeru mittsu no shishitsu, noryoku wo hakaru supikingu hyoka ruburikku no yuyosei [Usefulness of speaking assessment rubrics that assess three attributes and abilities, as defined in the next Course of Study]. *EIKEN BULLETIN*, *31*, 32–57. https://www.eiken.or.jp/center_for_research/list_1X/31/

Khabbazbashi, N., & Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. *Language Testing*. Advance online publication. https://doi.org/10.1177/0265532219898635

Kobayashi, S. (2005, June 25). *Holistic evaluation of EFL interview tests: Which is more reliable, evaluating holistically or analytically?* [Paper presentation]. 35th Annual Conference of the Chubu English Language Education Society, University of Yamanashi.

Koizumi, R., In'nami, Y., & Fukazawa, M. (2016). Multifaceted Rasch analysis of paired oral tasks for Japanese learners of English. In Q. Zhang (Ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 89–106). Springer Singapore. https://doi.org/10.1007/978-981-10-1687-5

Linacre, J. M. (2020a). *Facets: Many-facet Rasch measurement* (Version 3.83.1) [Computer software]. MESA Press. https://www.winsteps.com/facets.htm

Linacre, J. M. (2020b). *A user's guide to FACETS Rasch-model computer programs: Program manual 3.83.1*. http://www.winsteps.com/manuals.htm

Matsumura, K., & Moriya, R. (2019). Kyoshitsu niokeru Paired oral test no shindantekihyoka oyobi gakushusha no juyo nikansuru chosa: Kongo kenkyuho wo mochiite [Diagnostic assessment and learners' perceptions related to a classroom-based paired oral test: A mixed-methods approach]. *EIKEN BULLETIN*, *31*, 212–234. https://www.eiken.or.jp/center_for_research/list_1X/31/

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52–75. https://doi.org/10.1177/026553229000700105

McNamara, T. F. (1996). *Measuring second language performance*. Addison Wesley Longman.

McNamara, T., Knoch, T., & Fan, J. (2019). *Fairness, justice, and language assessment*.

Oxford University Press.
Metruk, R. (2018). Comparing holistic and analytic ways of scoring in the assessment of speaking skills. *Journal of Teaching English for Specific and Academic Purposes*, *6*(1), 179–189. https://doi.org/10.22190/JTESAP1801179M
Naito, T. (1995). Speech niokeru analytic evaluation to holistic evaluation [Analytic and holistic evaluation in speech assessment]. *STEP BULLETIN*, *7*, 146–151. https://www.eiken.or.jp/center_for_research/list_1/archives/
Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Peter Lang.
Negishi, J. (2011). Characteristics of group oral interactions performed by Japanese learners of English. (Publication No. 5722) [Doctoral dissertation, Waseda University]. Waseda University Repository. http://hdl.handle.net/2065/37662
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. https://doi.org/10.20581/arele.26.0_333
Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, *31*(2), 147–175. https://doi.org/10.1177/0265532213514401
Ono, M., Yamanishi, H., & Hijikata, Y. (2019). Holistic and analytic assessments of the TOEFL iBT® integrated writing task. *JLTA Journal*, *22*, 65–88. https://doi.org/10.20622/jltajournal.22.0_65
Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing*, *34*(4), 477–492. https://doi.org/10.1177/0265532217711431
Takaki, S. (2017). Kaikibunseki [Regression analysis]. In A. Hirai (Ed.), *Kyoiku shinri kenkyu notameno deta bunseki nyumon* [Introduction to data analysis for research in education and psychology] (2nd ed.). Tokyo Shoseki.
Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing*, *2*(1), 59–92. http://ijlt.ir/2019/07/16/vol-2-no-1-march-2012/
Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). *ETS Research Report Series*, *2006*(1), i–71. https://doi.org/10.1002/j.2333-8504.2006.tb02013.x
Zhang, X. (2019). *An investigation of the construct validity of the CET-SET: Perspectives on the construction and use of rating scales*. [Unpublished doctoral dissertation]. Shanghai Jiao Tong University.

Appendix

Appendix A: *Holistic Rubric*

3	Satisfies adequately 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.)
2	Satisfies to a certain degree 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	Needs more effort 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.

Appendix B: *Analytic Rubric With Four Criteria*

	Pronunciation & intonation	Grammar & Vocabulary		Fluency	Interactive communication
3	Japanese language interference of prosodic features and individual sounds are noticeable. However, constant attempts at assimilation/elision and to use appropriate rhythm make utterances reasonably easy to understand. Once or twice puts some strain on the listener and impedes understanding but not often.	Most basic structures[a] are sound. There are some inaccuracies, which however do not impede meanings, when complex structures are attempted[b]. Once or twice impedes communi-cation.	Generally, uses adequate range of vocabulary to manage most everyday topics, although experiences difficulty when required to expand on topics. Lack of vocabulary impedes communi-cation once or twice.	Hesitation while searching for language may be noticeable and speech may be slow, which, however, does not demand unreasonable patience of the listener. Once or twice demands unreasonable patience of the listener.	Communicates effectively by appropriately participating in turn-taking. Responds, comments (e.g., agree/disagree), asks questions, negotiates meanings verbally and non-verbally and develops the interaction in some but not all the occasions. Interaction is ineffective once or twice.
2	Japanese interference in prosodic features and individual sounds is marked. Some attempts at assimilation/elision and to use appropriate rhythm are shown. Occasionally puts some strain on the listener, but does not really impede understanding.	Basic structures[a] are occasionally inaccurate. Has just enough grammar to manage to get meaning across in everyday topics. More complex structures[b] are not attempted or not intelligible.	Choice of words is occasionally inaccurate in everyday-topics. Limitation of vocabulary may prevent discussion at some stages of the interaction[c], but does not really impede communi-cation.	Speech is slow and hesitant[d]. It occasionally demands unreasonable patience of the listener, but does not really impede communi-cation.	Communicates adequately in most everyday contexts, but could be rather passive with responding and commenting. Asks for clarification[e] verbally or non-verbally, although occasionally it may be unsuccessful. Not effective enough to contribute to develop the interaction.
1	*Speaks very frequently with mispronunciations and with Japanese katakana-like pronunciation/rhythm (without any assimilation/elision), which nearly always impedes understanding*[f].	Grammar is almost entirely inaccurate except for some stock phrases, which nearly always impedes communi-cation.	Shows only simplest words and phrases. Lack of vocabulary makes even basic communi-cation difficult.	Speech is very slow and disconnected. Almost impossible to follow, except for short or routine phrases.	Gives simple responses only when required, but is unable to maintain or develop the interaction. May show a few attempts[g] to ask for repetition or paraphrasing, which are nearly always unsuccessful.

Note. [a](e.g., phrases, simple/compound sentences). [b](e.g., complex sentence). [c](as he/she cannot express opinions properly). [d](e.g., with some unevenness and long pauses caused by rephrasing and searching for language). [e](repetition, paraphrasing). [f]This Level 1 description in Pronunciation was deleted and combined with Level 2. [g](mostly non-verbally).