Ogawa *Language Testing in Asia* (2022) 12:4
https://doi.org/10.1186/s40468-022-00154-9

Language Testing in Asia

RESEARCH Open Access

CAF indices and human ratings of oral performances in an opinion-based monologue task

Chie Ogawa

Correspondence: chieog@cc.kyoto-su.ac.jp
Faculty of Cultural Studies, Kyoto Sangyo University, Kyoto, Japan

Abstract

This study explored two assessment approaches to oral performances: analytical complexity, accuracy, and fluency (CAF) indices and human raters' evaluations. CAF indices are frequently used in second-language speaking (L2) research; however, because tasks are communicative and goal-oriented, the degree to which students achieve such communicative goals must also be included. By incorporating human ratings of monologue organization and perceived CAF into speaking assessments, researchers can better understand the relationship between the analytical CAF indices and human ratings of a monologue task. The participants consisted of 48 English as a Foreign Language (EFL) students in a Japanese university. Their oral performances of 2-min opinion-based monologues were audio-recorded and then transcribed and analyzed using CAF measures. In addition, 11 human raters evaluated the same recordings in terms of the following criteria: topic organization, complexity, accuracy, and fluency. These ratings were then analyzed using the many-facet Rasch measurement (MFRM). Multiple linear regression results showed that fluency accounted for a significant amount of the human ratings, but other measures (lexis, complexity, accuracy) explained only a small portion of the variance. This study concluded with implications regarding L2 testing in speaking assessments.

Keywords: Speaking assessment, CAF indices, Human ratings, Oral performances

Introduction

Second-language acquisition (SLA) scholars have been increasingly focusing on speaking performance assessment. Speaking tests are conducted in many different ways such as interview Q&As, role plays, and asking a test-taker to read a given text, state their opinions, solve a problem, describe a picture, or narrate a cartoon story. Some English proficiency tests (e.g., EIKEN, TEAP, TOEFL iBT, etc.) include opinion-based monologue tasks requiring examinees to discuss their preference or opinion about a certain topic with questions such as "Do you agree with the following statement: . . . ?" "Which is better, A or B?" or "What is your opinion about . . . ?" Oral performances are assessed primarily to evaluate the extent of a test-taker's ability to successfully convey meaning through speech or whether their speaking ability meets the minimum

requirements of the proficiency level. The following section discusses two methods of evaluating second-language (L2) oral performances: measuring recorded data using complexity, accuracy, and fluency (CAF) and human raters' oral performance.

Complexity, accuracy, and fluency indices

CAF indices have been routinely used as indicators of learners' oral proficiency and language acquisition (Housen et al., 2012). Skehan (1996) introduced a speaking proficiency model using the terms "complexity," "accuracy," and "fluency." These three CAF dimensions have been identified as distinct areas of L2 performance based on factor analyses (Housen et al., 2012; Norris & Ortega, 2009). One advantage of using the CAF measures is that doing so allows researchers to capture L2 learners' performance and proficiency comprehensibility. Most task-based studies have employed analytical measures for objective evaluations of transcribed speech data using CAF indices, because L2 performance has multiple components as conveyed by the concepts of complexity, accuracy, and fluency.

Syntactic complexity in speech research has been generally measured using speech units such as the analysis-of-speech unit (AS-unit; Foster et al., 2000) and by calculating sentence length (e.g., mean length of AS-units; Norris & Ortega, 2009), with a longer AS-unit indicating a speaker's ability to produce more complex utterances. Another way to assess syntactic complexity is based on subordination, which refers to the number of clauses in an AS-unit (Norris & Ortega, 2009); the higher the number of subordinate clauses in an AS-unit, the more complex the utterance. The second part of the CAF framework, accuracy, refers to the target language. Housen et al. (2012) expanded their definition of accuracy to include appropriateness and acceptability, as the norm changes depending on the social context.

Meanwhile, fluency refers to the smooth and speedy delivery of speech without pauses, repetitions, or repairs (De Jong et al., 2015). Three subdimensions are recognized in utterance fluency: speed fluency, breakdown fluency, and repair fluency (Tavakoli & Skehan, 2005). Speed fluency, which is the speed or density of linguistic units, is usually measured via speech rate (e.g., number of syllables per minute). Breakdown fluency refers to the number, length, and location of pauses (Housen et al., 2012) and is usually identified by measuring pauses or identifying where they occur in an utterance (e.g., pauses at the end or in the middle of a clause). Repair fluency pertains to the frequency of speakers' false starts, self-corrections, or repetitions. Yan et al. (2021) found that micro-level fluency features (e.g., mean length of run and location of pauses) had stronger explanatory power in terms of indicating how speech is perceived than macro-level features (e.g., the amount or rate of speech production). Compared to the measurable aspects of utterance fluency, perceived fluency describes the listener's impression of the speaker's fluency (Segalowitz, 2010). In this regard, perceived fluency is the perception of how easily and efficiently the listener can comprehend the speech. Perceived fluency involves human raters, which will be explained in the following section.

Human ratings

Speech research (e.g., De Jong & Vercellotti, 2016; Tavakoli & Skehan, 2005) has extensively used analytical CAF measures, but certain issues have been reported with these

measures. First, because they involve time-consuming procedures, analytical CAF measures are rarely used in classroom assessment. These procedures include transcripts of recorded speech, AS-unit-based sorting, calculations of morphosyntactic errors, number of clauses, syllables, pause length, and pause location.

Second, analytical CAF measures do not provide insights into students' extent of communicative achievement. Although these quantifiable CAF measures have been widely used in SLA speaking research, they do not reliably indicate the degree to which L2 speakers achieve communicative goals (Kuiken & Vedder, 2017; Pallotti, 2009).

Third, analytical CAF measures might fail to provide sufficient ecological validity. When performing CAF analysis, researchers usually interpret the results as "the more, the better." However, faster speech cannot always guarantee a better understanding from the listeners' point of view. In this regard, perceived fluency (Segalowitz, 2010) or comprehensibility (Suzuki & Kormos, 2020; Saito, 2021), which emphasizes the listener's point of view, must be considered alongside the analytical measures.

For these reasons, solely relying on analytical CAF measures is insufficient when assessing oral performance (De Jong et al., 2012; Ortega, 2003; Kuiken & Vedder, 2017; Pallotti, 2009). To better evaluate learners' speaking performances and to improve the use of such an approach in real-world settings, human raters should be employed to complement analytical CAF measurements (e.g., De Jong et al., 2012; Iwashita et al., 2008; McDonald, 2018; Magne et al., 2019; Suzuki & Kormos, 2020; Tran & Saito, 2021).

Raters play important roles in assessing examinees' language proficiency. Specifically, the communicative component in language testing is considered essential. Several studies have added human raters' holistic judgments to analytical CAF indices to evaluate L2 speakers' task-based performances (e.g., Revesz et al., 2016; Suzuki & Kormos, 2020). Holistic assessment in these studies often have slightly different emphases, such as the overall judgment of a speaker's degree of comprehensibility (Suzuki & Kormos, 2020; Saito, 2021), communicative adequacy (Revesz et al., 2016), overall communicative effectiveness (Sato, 2012), and communication ability (Sato & McNamara, 2019), or proficiency level (Yan et al., 2021), but the findings of previous studies show some features of human raters' impression of L2 speakers' oral performances.

First, perceived fluency tends to strongly predict holistic ratings. Suzuki and Kormos (2020) found a strong association between perceived fluency and holistic rating, in which raters intuitively judged comprehensibility using a nine-point Likert scale. In Sato's (2012) study, among grammatical accuracy, fluency, vocabulary range, pronunciation, and content elaboration/development, fluency was the second strongest predictor of overall effectiveness, followed by content development. Both studies (Suzuki & Kormos, 2020; Sato, 2012) required raters to intuitively assign holistic scores without detailed descriptions. Previous findings indicate that the degree to which a speaker sounds fluent from a listener's point of view is important for communicative success.

Second, researchers have also found a relationship between human raters' holistic judgment of oral performances and analytical fluency measures. According to Revesz et al. (2016), a set of linguistic factors significantly influenced holistic communicative adequacy as perceived by trained raters. The frequency of filled pauses (breakdown fluency) was the strongest predictor, with fluency emerging as a critical determinant of

holistic oral performance ratings. Suzuki and Kormos (2020) also found that speed fluency measures (articulatory speed of individual words) strongly influenced ratings of perceived comprehensibility.

Third, nonlinguistic components, such as speech organization and speech elaboration, also predict raters' holistic ratings strongly. Sato's (2012) standardized regression coefficients showed that content elaboration/development was the strongest predictor and a crucial component of oral performance as opposed to other linguistic features including grammatical accuracy, fluency, vocabulary range, and pronunciation. Indeed, in another study by Sato and McNamara (2019), findings from interviews and stimulated recalls showed that linguistic correctness was not necessarily the main point of raters' evaluations of communicative effectiveness, but raters positively assessed a speaker who successfully completed a task or provided better-quality content.

It would follow that intuitive holistic judgments using raw scores can be meaningful depending on the research purpose, and utilizing the holistic judgment might have strong ecological validity, as it might reflect the impressionistic judgment made by the listener during real-life communication (Saito, 2021). However, it is still not clear that what kinds of constructs the holistic rating possesses. In addition, these intuitive ratings can also be affected by factors related to listeners (Saito, 2021). For example, even if two listeners assess the same speech, their ratings may differ. In this regard, the many-facet Rasch measurement (MFRM) is useful in analyzing performance data that involves three or more components such as test-takers, raters, and the evaluation criteria (Linacre, 2002). The MFRM allows for the inclusion of additional performance test variables as *facets* and an assessment of participants' performances based on several such facets in the performance setting. This measurement approach provides a breadth of information of how raters empirically judge participants' oral performances. Although the application of Rasch measurement in language assessment has gradually increased (Aryadoust et al., 2021), more studies would be needed to provide more detailed results regarding the MFRM's role in analyzing speaking performances.

This study seeks to fill the literature gap on speaking task performance assessment based on CAF indices and intuitive human ratings and provide more evidence from MFRM data. This study specifically examines the following research questions:

1. How do analytical rating scales based on organization and CAF evaluate opinion-based monologue tasks?
2. What do analytical CAF measures contribute to human ratings of the same opinion-based monologue tasks?

Method

Participants and context of the study

Forty-eight first-year Japanese students attending a private university in eastern Japan participated in the study. Eighteen of them were male and 30 were female, with an average age of 18.08 years (SD = .27). The participants' proficiency levels were from low-intermediate to intermediate (TOEIC range of 350–550). The author was the teacher of these classes. All participants were informed of the purpose of the study, and they signed the consent form.

Data collection procedure

Data collection was conducted in intact classes, in which the author taught. The participants performed 2-min opinion-based monologues that were recorded three times in one academic semester (during weeks 2, 8, and 14). Before the monologues were recorded, the participants were given 1 min of planning time. They were asked to write their ideas on a blank paper, which were then collected before recording so that they would not refer to any materials while speaking. They produced a total of 144 (48 participants × 3 times) recordings. Appendix A shows the questions.

Design of the study

This study is a part of the larger study and mostly employs a quantitative research design. However, qualitative data (speech transcription) were used in order to follow up the quantitative results. The researcher transcribed the recorded data including fillers and self-repetitions. At this time, pause length was not included. A total of 288 min of speech data were transcribed (2 min × 48 participants × 3 times). The transcriptions were then double-checked by a research assistant. The original transcription was used when transcribing pruned speech, marking AS-unit boundaries and clauses and measuring pauses using the Praat speech analysis software.

After the speech samples were transcribed, transcriptions of pruned speech, which excluded fillers, self-corrections, and repetitions, were produced to examine syntactic complexity and accuracy. Pruned speech was used for assessing syntactic complexity to avoid incorrectly measuring complex sentences. Pruned speech was also used to calculate syntactic accuracy so that self-corrections are accounted for after the speakers noticed syntactic errors. For example, if a speaker made a self-correction such as "She {weared} . . . was wearing," it was accepted as a correct utterance because the speaker noticed the error and self-corrected; pruned speech avoids the possibility of decreasing syntactic accuracy measures.

AS-units

AS-units were used to evaluate syntactic complexity and morphosyntactic accuracy (e.g., De Jong & Vercellotti, 2016; Foster et al., 2000; Lambert et al., 2017; Nitta & Nakatsuhara, 2014; Revesz et al., 2016; Tavakoli et al., 2016). Foster et al. (2000) stated that AS-units effectively capture aspects of spoken language that other units might miss or categorize as errors.

Syntactic complexity

Syntactic complexity was measured using (a) the mean clause length (pruned speech) (numbers of words/AS-unit) and (b) clauses per AS-unit (pruned speech). When calculating both complexity measurements, pruned speech was used. For clauses per AS-unit, the subordination figure was calculated by counting all clauses and dividing them by the number of AS-units.

Lexical diversity

This study used the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010) (http://textinspector.com/workflow), which can assess lexical diversity without the influence of text length (McCarthy & Jarvis, 2010).

Accuracy

Global accuracy (morphosyntactic and lexical accuracy) was evaluated after pruning. Morphosyntactic and lexical accuracy refers to one's ability to avoid morphosyntactic errors (Ellis, 2009; Skehan & Foster, 1999), which can occur with inflectional morphemes (e.g., third-person singular *-s*, plural *-s*), function words (e.g., articles, prepositions), content words (e.g., adjective–noun collocations), and Japanese use (e.g., *igirisu* for England). When an utterance made no sense, however, the error type could not be determined.

The researcher calculated the error-free AS-units as follows. First, the errors in the transcription were counted based on the above criteria. Then, the total number of AS-units and the number of error-free AS-units were counted for each recorded monologue. The ratio of error-free AS-units for each speech was calculated by dividing the number of error-free AS-units by the total number of AS-units.

Fluency

Mean pause length Pauses are classified into silent and filled. Silent pauses were defined as those longer than 300 ms (e.g., De Jong & Bosker, 2013; Thai & Boers, 2016), and nonverbal fillers such as *uh*, *ah*, and *um* were also treated as pauses (e.g., De Jong & Perfetti, 2011; De Jong et al., 2013; De Jong et al., 2015). Hence, in this study, *pauses* included both types. Pause length was measured using Praat (http://www.praat.org; Boersma & Weenink, 2009).

Number of repairs Repair fluency includes false starts, reformulations, and repetitions of words or phrases (Tavakoli & Skehan, 2005, p. 255). In this study, all three were counted as repairs. Fillers were not considered part of repair fluency because they were already included in breakdown fluency.

Mean length of run The mean run length was calculated as the mean number of syllables produced in an utterance between pauses (total syllable count divided by run count). A run is a fluent sequence between two silent pauses. Run count was calculated by adding 1 to the number of pauses; for example, if there were seven pauses, then there were eight runs (7 + 1 = 8), and the total syllable count would be divided by 8. Syllables were counted using the Syllable Count website (Arczis Web Technologies, 2019).

Phonation time ratio The phonation time ratio was calculated as the total length of phonation time (time spent speaking) divided by the total response time a participant spent speaking (2 min). To calculate the measure, first, the total length and number of pauses were determined using a cut-off rate of 300 ms. Phonation time was determined by subtracting the total time of silent pauses from the total response time (e.g., 120-s total − 30-s pause length = 90 s).

Mean duration of syllable Speed fluency was calculated as the average duration of syllables, that is, speaking time divided by the number of syllables produced (e.g., Bosker

et al., 2013; De Jong et al., 2013, 2015). This allows speech fluency to be separated from other disfluency components such as pauses and repairs, which are unconfounded (De Jong et al., 2015). The mean syllable duration was analyzed using speaking time after excluding pauses.

Table 1 shows the calculations for the CAF measurements in this study. Five fluency measures were included: mean length of pauses, number of repairs, mean length of syllable, mean length of run, and phonation time ratio. Pauses include both silent and filled pauses.

Intercoder reliability

Two raters performed CAF analysis on the transcribed data. First, a research assistant double-checked all transcriptions. Second, the researcher coded all the data. To ensure the reliability of the CAF measures, approximately 10% of the total sample size were also calculated by a research assistant. Percentage agreements were determined for the classification of student output into AS-units and clauses. Initially, the percentage agreement was 73.3% for AS-units, 86.6% for clauses, and 80.0% for error-free AS-units. All coded transcripts were compared, disagreements discussed, and agreements reached for every case. The data were then rechecked, and intercoder agreement was found to be 100%. Word count and syllable count were computed using website software, so intercoder reliability was not calculated for these aspects.

Human ratings

Human ratings were employed alongside the analytical CAF measures. The goal of the opinion monologue task was for speakers to successfully perform a coherent, organized monologue with sufficient information. Therefore, in this study, human raters assessed both linguistic competence and topic organization to achieve the task's communicative goal.

First, the rating scale for organization was developed based on the idea that a coherent, well-organized speech would allow listeners to clearly understand the message. The researcher developed the rating criteria to reflect the need for a descriptor for each point to match the level of difficulty of descriptors across all four rating scales (e.g., McDonald, 2018). Second, the rating scales for the CAF criteria were adapted from

Table 1 CAF measurements

	Type	Specific measure	Calculation
Complexity	Syntactic complexity	Clauses per AS-unit	Number of clauses/number of AS-units
		Mean length of AS-units	Number of words/number of AS-units
	Lexical diversity	MTLD	MTLD
Accuracy	Global measures	% of error-free AS-units	Number of error-free AS-units/total number of AS-units
Fluency	Breakdown fluency	Mean length of pauses	Sum of pauses/number of pauses
	Repair fluency	Number of repairs	Total number of repairs
	Speed fluency	Mean length of syllables	Spoken time/number of syllables
	Combination	Mean length of run	Total number of syllables/number of runs
		Phonation time ratio	Spoken time/total time

Note. Spoken time is phonation time spent speaking without silent pauses and fillers and includes repairs

Iwashita et al. (2001) and modified accordingly. A five-point scale was used to decrease the raters' cognitive load (Nemoto & Beglar, 2014): 1 = *Unsuccessful performance*, 2 = *Poor performance*, 3 = *Moderately successful performance*, 4 = *Successful performance*, 5 = *Very successful performance*. Table 2 shows the final rubric consisting of four rating scales assessed along these five levels.

After the rubric was developed, 10 additional raters were recruited while the researcher acted as the 11th rater. All raters were university English teachers and held master's degrees in applied linguistics or related fields. Six raters were Japanese, one Canadian, one British, one Australian, and one Chinese. All of them underwent rater training for approximately 40 min to allow them to understand the criteria for evaluating each component—organization, complexity, accuracy, and fluency—and the general evaluation standard. First, the researcher explained the rating tasks and the rubric. The raters then listened to four sample performances and assessed them using a handout (Appendix B). Each sample audio file was from a different experimental group and a different test time. Next, the raters and the researcher discussed their ratings and the reasons for them. After the training, they rated 20–40 speeches at their own pace at home, while the researcher evaluated all 144 samples.

Analysis

The facets in this study were person ability, rater severity, and rating category difficulty. Person ability was estimated while considering the effects of the other facets. The logit person ability measures were produced from the FACETS analysis, which represents a single combined measure of the scores from the four rating scales considering the effects of other facets such as rater strictness and scale difficulty.

First, the raters' raw scores were statistically analyzed using the MFRM in FACETS version 3.71.4 (Linacre, 2013). A total of 144 distinct participant codes were considered

Table 2 Human raters' rubric

	Organization	Complexity	Accuracy	Fluency
5 Very successful	Speech is extremely well organized and very coherent with detailed information.	A wide range of grammar is used. Attempts to use coordination and subordination to convey ideas very often.	Grammatical errors are absent or very rare.	Speech is extremely smooth. Hesitations rarely occur, and they are very short.
4 Successful	Speech is fairly well organized and coherent.	A fairly wide range of grammar is used. Often attempts to use coordination and subordination to convey ideas.	Grammatical errors are rare.	Speech is fairly smooth. Hesitations very occasionally occur, and they are short.
3 Moderately successful	Speech is somewhat well organized and mostly coherent.	A somewhat wide range of grammar is used. Occasionally attempts to use coordination and subordination to convey ideas.	Grammatical errors sometimes occur.	Speech is somewhat smooth. Hesitations occur occasionally, but they are sometimes lengthy.
2 Poor	Speech is not well organized and is somewhat incoherent.	A limited range of grammar is used. Mostly relies on single clauses and simple phrases.	Grammatical errors often occur.	Speech is disfluent. Hesitations are frequent and often lengthy.
1 Unsuccessful	Speech is very poorly organized and is incoherent.	An extremely limited range of grammar is used. Completely relies on single clauses and simple phrases.	Grammatical errors very often occur.	Speech is very disfluent. Hesitations are very frequent and sometimes very lengthy.

for MFRM analysis. Second, Pearson correlation and multiple regression analyses were performed using the logit person ability measures produced from the FACETS analysis.

Results

The FACETS results initially showed two misfit raters: rater 3 was overly restrictive (infit MNSQ = .63), while rater 4 was overly erratic (infit MNSQ = 1.91). Therefore, they were excluded from the analysis. Afterward, 11 recordings were only single-rated in this dataset.

Table 3 shows the raters' Rasch statistics for the monologue task. The results of the FACETS analysis indicated mean Rasch difficulty estimates (measure) ranging from −1.05 to 1.05 for the nine raters. Rater 7 had the highest severity estimate followed by raters 1, 6, 10, 5, 9, 11, 8, and 2 (Table 3); this meant that rater 7 was the most restrictive evaluator, while rater 2 was the most lenient. Rater infit and outfit were acceptable for all remaining raters; that is, they were not erratic or overly restrictive in their use of the scales (infit MNSQs between .78 and 1.29). Rater reliability (.93) was high, while separation (3.56) was moderate.

The scores for each criterion were analyzed to examine interdependent patterns in the criteria. The mean Rasch item difficulty estimates for each rating component ranged from -.44 to .76 (Table 4). Fluency (.76 logits) had the highest difficulty estimate followed by organization, complexity, accuracy (− .14, − .17, − .44 logits, respectively); thus, fluency was the most difficult criterion, while accuracy was the easiest criterion on which to achieve a high score. Complexity and organization were equally easy. All items—fluency, complexity, organization, and accuracy—met the infit MNSQ criterion of .50–1.50 (Linacre, 2002).

A Rasch principal component analysis (PCA) of item residuals analysis was conducted to determine the dimensionality of the rating scales. The Rasch model explained 61.4% of the variance (eigenvalue = 6.35) and unexplained variance in the first contrast = 1.5 (14.3%). They generally met Linacre's (2017) requirements that over 50% of the variance be explained by the Rasch measures and that the largest secondary dimension have an eigenvalue less than 2.0. However, the first contrast did explain over 10% of the total variance. These results indicated that the monologue measure was fundamentally unidimensional, which refers to the assumption that the test measures only one underlying latent trait (Aryadoust et al., 2021).

Table 3 Rasch statistics for the raters for the speaking assessment

Rater	Measure	*SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
7	1.05	0.18	0.94	− 0.35	0.94	− 0.35	0.73
1	0.38	0.18	0.90	− 0.64	0.91	− 0.56	0.59
6	0.24	0.18	0.97	− 0.14	0.98	− 0.08	0.72
10	0.18	0.07	1.08	1.37	1.08	1.32	0.64
5	0.07	0.13	0.81	− 1.88	0.81	− 1.91	0.83
9	− 0.02	0.19	1.29	1.73	1.3	1.75	0.59
11	− 0.14	0.15	0.99	− 0.01	0.99	− 0.04	0.67
8	− 0.71	0.18	0.78	− 1.51	0.78	− 1.48	0.56
2	− 1.05	0.19	0.98	− 0.1	0.97	− 0.15	0.72

Table 4 Rasch statistics for human rating components

Rating criterion	Measure	*SE*	Infit MNSQ	Outfit MNSQ	Pt-measure correlation
Fluency	0.76	.09	0.98	0.98	.73
Organization	− 0.14	.09	1.16	1.15	.73
Complexity	− 0.17	.09	0.68	0.68	.73
Accuracy	− 0.44	.09	1.19	1.18	.57

For model fit, 62 of the 1,336 valid responses modeled (4.6%) were found to be associated with standardized residuals greater than or equal to 2.0, while three responses (0.002%) were found to be associated with standardized residuals greater than or equal to 3.0. This also meets Linacre's (2017) model-fit stipulations that less than about 5% be greater than or equal to 2.0 and about 1% or less be greater than or equal to 3.0.

The FACETS map in Fig. 1 provides an overview of the rating results. All facets were measured in uniform units (log-odds = *logits*) indicated on the left side of the map in the *Measure* column. The second column, *Participant*, which represents three separate time points, shows the participants' Rasch ability estimates. The more proficient participants are placed toward the top and the less proficient ones toward the bottom. The third column, *Raters*, shows the rater severity estimates. The more severe raters appear toward the top, while the more lenient ones appear toward the bottom. The fourth column, *Ratings*, shows the difficulty levels of the four rating categories: fluency was the most difficult, followed by organization, complexity, and accuracy. The last column, *Scale*, shows the category thresholds separating the different scoring levels along the combined five-point scale.

Table 5 shows summary statistics for the MFRM analysis of the combined rating scale showing person reliability (.85) and person separation (2.39).

Table 6 shows the Rasch rating category statistics for human ratings for the monologue performances. All categories functioned well according to the diagnostic criteria: category frequency, average measures, threshold estimates, category fit, and probability curves. Almost 40% of participants' speeches were rated as moderately successful (130 counts, 39%), 10% (34 counts) were rated as very successful, and 1% (5 counts) were rated as unsuccessful. However, the results failed to meet one of Linacre's (2002) guidelines for evaluating rating scale category functioning, as there were less than 10 observations at scoring level 1.

Table 7 shows the descriptive statistics for the CAF measures and human ratings. Pearson correlation was used to understand the interrelations among predictor variables (Table 6). According to Plonsky and Oswald (2014), *r* values close to .25 indicate a small effect, .40 a medium effect, and .60 a large effect in L2 research. A high correlation was observed between two complexity measures (r = .79) and between mean length of run and phonation time ratio (r = .66). However, none of the variables had a correlation coefficient of above .90 (Tabachnick & Fidell, 2007, p. 90), which indicates multicollinearity.

Pearson correlation was also used to understand the interrelation among predictor variables (Table 8). A high correlation was found between two complexity measures (r = .79), and mean length of run and phonation time ratio were highly correlated as well (r = .66).

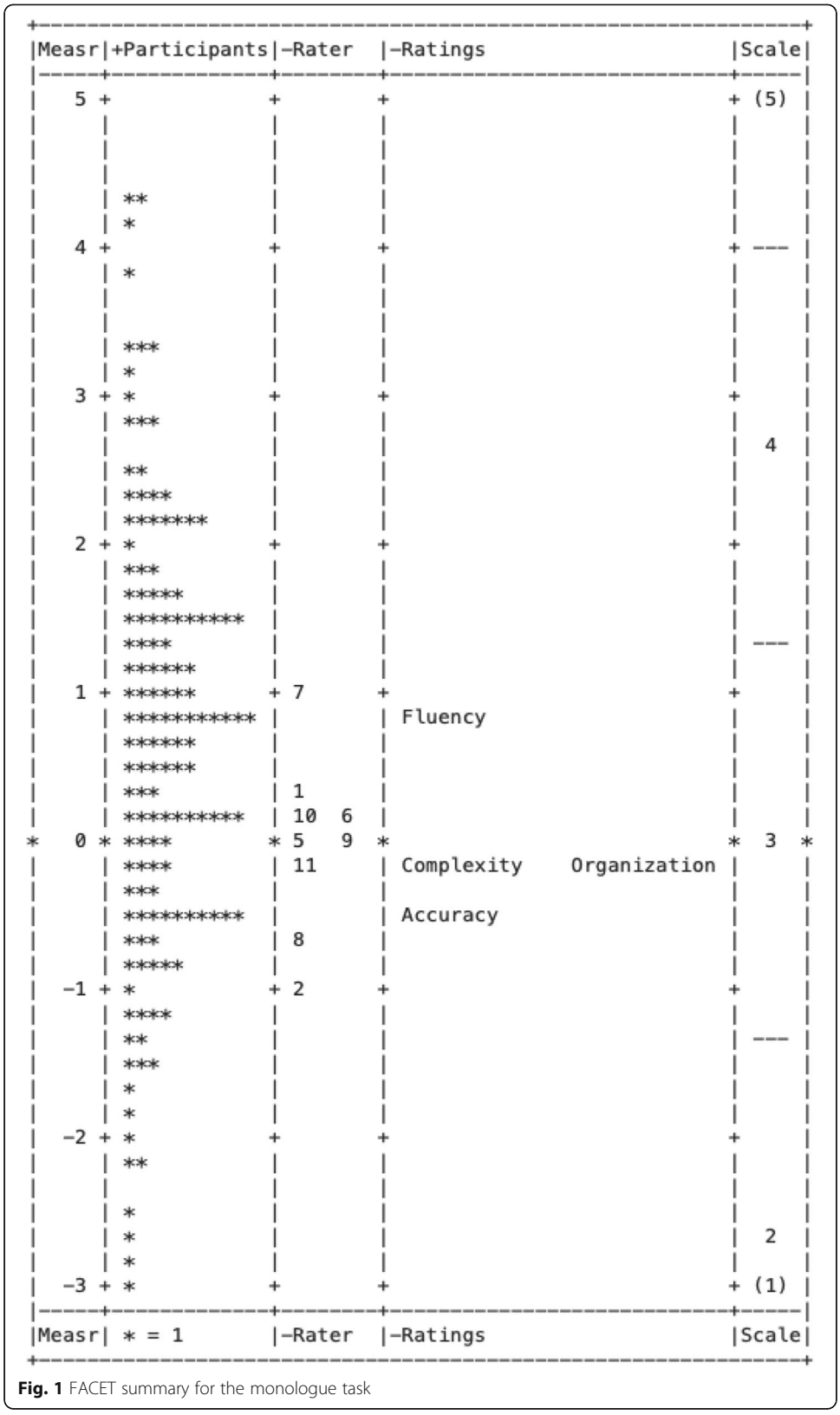

Fig. 1 FACET summary for the monologue task

Table 5 Summary statistics for the MFRM analysis of the combined rating scale

Statistics	Participants	Raters	Criteria
M measure	0.63	0.00	0.00
M SE	0.56	0.16	0.09
χ^2 (fixed)	1009.3*	92.2*	103.5*
df	143	8	3
Separation ratio (sample)	2.39	3.56	5.79
Separation reliability (sample)	.85	.93	.97

Note. *$p < .01$

To determine the relative weights of CAF measures in human ratings for oral performances, hierarchical multiple linear regression analysis was conducted. The human raters' scores using FACET measures were chosen as the dependent (predicted) variables, and analytical CAF measures were selected as the predictor variables. The assumption of multiple regression was reviewed, and its requirements were met. Based on previous studies (e.g., Revesz et al., 2016), it was hypothesized that fluency is a strong predictor followed by complexity, lexis, and accuracy. Therefore, the fluency scores (mean syllable length, mean pause length, repair, mean run length, phonation time ratio) were entered for the first step, while the complexity scores (mean AS-unit length and clauses per AS-unit) were entered for the second step. Lexis was entered for the third step, while accuracy scores were entered for the last step. The first step showed that fluency accounted for a significant amount of raters' assessment (R^2 = .43, $F(5, 138) = 20.78$, $p < .001$). Lexis was the next strongest predictor (R^2 = .041, $F(1, 137) = 20.25$, $p < .001$). Complexity (R^2 = .017, $F(2, 135) = 16.05$, $p < .001$) and accuracy measures (R^2 = .016, $F(1, 134) = 15.72$, $p < .001$) accounted for raters' assessment as well. Table 9 shows the multiple regression analysis results.

According to the model, human raters' perceptions of speaking performances were primarily predicted by analytical fluency measures. Meanwhile, other measures (complexity, lexis, accuracy) were not likely to influence raters' assessment.

Because fluency plays a major role in human raters' evaluation of oral performances, each fluency measure must be examined more in detail. Table 10 reports the degree to which each fluency variable in the model contributes to the prediction of human raters' assessment. The standardized regression coefficients indicate that the strongest predictor was phonation time ratio (β = .63) followed by repairs (β = − .37) and syllable length (β = − .16). These suggest that the length of speaking time was the strongest driver of human raters' judgment of oral performances.

Table 6 Ratio of human rating categories

Category	Observed count	Average measure	Outfit MNSQ	Threshold calibration	Threshold change
1 unsuccessful	5 (1%)	− 2.09	1.00	–	–
2 poor	72 (22%)	− .74	1.10	− 4.13	–
3 moderately successful	130 (39%)	.63	.90	− .61	+ 3.52
4 successful	93 (28%)	1.66	1.00	1.52	+ 2.13
5 very successful	34 (10%)	2.94	.90	3.22	+ 1.70

Note. All of Linacre's (2002) guidelines were met

Table 7 Summary statistics of human ratings and CAF measures

	Mean	SD	Minimum	Maximum
Human ratings (FACET measure)	0.63	1.47	− 2.95	4.29
Complexity				
Clauses/AS	1.71	0.35	1.07	3.00
ML of AS-unit	10.76	2.09	6.67	17.00
Lexical diversity	37.14	10.38	18.19	64.03
Accuracy				
% of error-free AS-unit	0.64	0.18	0.11	1.00
Fluency				
MDS	.305	.039	.217	.421
ML pauses	0.91	0.42	0.30	2.58
Repair	9.81	5.96	0.00	35.00
MLR	4.65	0.96	2.61	7.23
PTR	51.94	9.09	32.90	70.70

Note. ML mean length, *MDS* mean duration of syllable, *MLR* mean length of run, *Ptr* phonation time ratio

Discussion

Raters' assessment of oral performances

To address research question 1 (How do analytic rating scales based on organization and CAF evaluate opinion-based monologue tasks?), the MFRM provided insights into how human raters perceived Japanese university students' oral performances in opinion-based tasks. The FACET analysis results confirmed unidimensionality. Although the participants had similar proficiency levels (TOEIC range of 350–550), the raters were able to consistently spread out their oral performances on the logit scale.

Table 8 Correlations among human ratings and CAF measures

	Human ratings	Clauses/AS	ML AS-unit	Lexical diversity	% error-free	Pause	Repair	MDS	MLR	PTR
Human ratings	–									
Complexity										
Clauses/AS	0.32**	–								
ML AS-unit	0.30**	0.79**	–							
Lexis										
Lexical diversity	0.31**	0.15*	0.27**	–						
Accuracy										
% error-free	0.18*	− 0.21**	− 0.23**	0.18*	–					
Fluency										
Pauses	0.25**	0.21*	0.17*	0.31**	0.03	–				
Repair	− 0.05	0.00	0.03	− 0.10	0.01	0.06	–			
MDS	− 0.13*	− 0.07	− 0.12	− 0.16*	0.01	− 0.05	− 0.19*	–		
MLR	0.46**	0.14*	0.22**	0.14*	0.10	0.30**	0.25**	− 0.32**	–	
PTR	0.53**	0.26**	0.32**	0.03	− 0.01	0.20**	0.43**	− 0.03	0.66**	–

Note. **Correlation is significant at < .01 (two-tailed). *Correlation is significant at < .05 (two-tailed). Human ratings = FACET measure. *MLR* mean length of run, *PTR* phonation time ratio

Table 9 Multiple regression analysis results using analytical CAF as predictors of human raters' evaluation

Predicted variable	Predictor variable	R^2	R^2 change	*F*	*p*
Human raters' scores	Fluency	.43	.430	20.78	$p < .001$
	Complexity	.47	.041	20.25	$p < .001$
	Lexis	.49	.017	16.05	$p < .001$
	Accuracy	.51	.026	15.72	$p < .001$

According to the FACET summary (Fig. 1), fluency had the highest difficulty estimate followed by organization, complexity, and accuracy; thus, fluency was the most difficult criterion on which to obtain a high score, while accuracy was the easiest. The raters were likely to evaluate fluency more strictly and accuracy more leniently. One reason why raters were strict about the fluency component is that they might have found that speaking smoothly to convey one's message was a salient feature in opinion-based tasks. As opposed to closed tasks such as picture descriptions, in which speakers must use expected grammatical and lexical items to describe a given image, opinion-based tasks allow speakers to express their ideas more freely (e.g., Suzuki & Kormos, 2020). In this type of flexible open tasks, raters judge more critically the fluent communication of one's message than their production of accurate utterances.

Another reason is that raters might have held higher standards for fluent performance than for other linguistic features such as accuracy and complexity and therefore assessed fluency more strictly to achieve high scores on. Because they were expert English teachers who taught communication courses in university, their backgrounds might have influenced them to be more stringent in evaluating how smoothly a speech was delivered and more tolerant toward morphosyntactic errors.

Besides severity, as shown in Table 8, fluency, complexity, and organization showed relatively high correlations with human ratings (r = .70, r = .73, and r = .71, respectively). That is, the higher the score of a component, the more likely it is for human ratings to be higher. Meanwhile, the part-measure correlation for accuracy was smaller (r = .59). Accuracy did not correlate with the raters' scores as much as the other components. Such a smaller coefficient implies that accuracy was a somewhat different criterion in the assessment of an opinion-based speaking task and indicates that human raters might have judged the participants' accuracy as a separate component from the other criteria.

Relative contribution of analytical CAF measures to human ratings

To address research question 2 (What do analytical CAF measures contribute to human ratings of the same opinion-based monologue tasks?), the results of multiple

Table 10 Multiple regression coefficients

Predictors	Standardized beta coefficients	Sig
(Constant) phonation time ratio	.63	.000
Mean length of run	.05	.645
Mean duration of syllable	− .16	.025
Mean length of pauses	.13	.061
Repair	− .09	.000

regression analysis provided a more detailed understanding of CAF measures and human ratings. The results showed that analytical fluency measures accounted for a significant amount of human rater evaluation (53% of variance), but the other analytical measures (lexis, complexity, and accuracy) explained only a small portion of the variances (1.5–3.4%). According to a previous study (e.g., Revesz et al., 2016), analytical fluency measures contribute significantly to human ratings of oral performances; the present study also found that among CAF measures, fluency is the most influential factor in human ratings.

Among the five fluency measures, phonation time ratio had the highest standardized beta coefficient (β = .74), indicating that the length of speaking time positively influenced human raters' evaluations with a large effect size. Phonation time ratio captures the proportion of the total utterance length to the total speech length produced. Fewer pauses usually generate an increase in phonation time ratio, as more time is spent speaking, and less time is spent pausing (Towell et al., 1996). This is reasonable because participants who spent time on speaking longer might express their opinions more in detail and convey meaning more successfully. In addition, those who spent more time speaking may produce more complex sentences. Indeed, the Pearson correlation coefficients (Table 8) show a positive relationship between phonation time ratio and clauses per AS-unit (r = .26) and between phonation time ratio and mean length of AS-unit (r = .32).

Repairs had the second most influential standardized beta coefficient (β = − .39) toward human ratings, suggesting that the amount of repairs negatively affected human ratings. The following are excerpts from a participant who received an extremely low score from the raters (logit = − 3.41):

(1.40) {i} i think club activity is (0.42) good idea for students (1.91)

(eh) {club activity} (2.25) (eh) {because (2.39) (eh) club active} (2.00) (eh) {making} (2.91) (eh) {make (4.12) (eh) club activities} (6.83) (eh) {make friends} (0.85) (eh) {make friends} (3.04) (eh) {make} (4.19) (eh) when join (0.43) club activity (0.48) make many friends (4.03)

While this participant clearly expressed their opinion in the first line, they reformulated their utterance and repeated "club activities" and "make friends" many times. Perhaps this participant was trying to think of what to say and how to say it at the same time and actually intended to articulate that "club activity is a good idea because you can make friends when joining a club." However, from the raters' perspective, excessive repetitions might have prevented them from understanding the participant's opinion. According to Revesz et al. (2016), repairs were more noticeable for high-proficient speakers than for low-proficient ones, indicating that repair frequency significantly affected human raters' evaluation of speech performance. This study shows that repairs might also negatively influence human ratings although the participants were low- to mid-proficiency students. Although Yan et al. (2021) explained that micro-level fluency (mean length of run, juncture pause rate, and repair success rate) are the key components to explaining L2 speakers' proficiency, this study suggests that macro-level features (such as counting the spoken time and repairs) remained beneficial indicators from the perspective of listeners.

Complexity and lexical diversity were the next influential predictors after fluency. Despite the miniscule variance, complex utterances and the wider variety of vocabulary positively affect human raters' assessment. Supporting Revesz et al.'s (2016) finding that raters appeared to rely on a range of vocabulary information (e.g., diversity) during L2

communicative adequacy judgments, this study also showed that lexical diversity was to some extent salient to the assessment of speakers' expression of their opinions.

Accuracy was the least influential predictor of human rating. The present findings suggested that error-free AS-units were not significantly important in obtaining high scores from human raters. This might be because raters were either unable to critically detect error-free AS-units or did not highly prioritize morphosyntactic accuracy. Indeed, as shown in the discussion of research question 1, the raters evaluated perceived accuracy in a lenient manner, implying that they may have been more tolerant of the students' syntactic errors.

From a listener's point of view, conveying messages was considered more important for success in real world communication. The raters in this study held some generous attitudes toward L2 learners' grammatical errors and evaluated the extent to which their messages were expressed. Similar findings were found involving different types of raters. Nonexperts in the English teaching field including both English native speakers and non-native speakers prioritized fluency over accuracy (e.g., Sato & McNamara, 2019; Suzuki & Kormos, 2020). Suzuki and Kormos (2020) explained that morphosyntactic errors had a weak association with raters' perceived comprehensibility. Therefore, this study supports the value of fluency in the expression of one's opinion regardless of listeners' L1 backgrounds or real world teaching expertise.

Several limitations may have affected the results of this study; hence, some results must be treated with caution. First, the results of the Rasch PCA of item residuals analysis showed an unexplained variance in the first contrast of 14.3%, which failed to meet Linacre's criteria (2017). We must consider that other variances can be explained for oral performance assessment alongside human ratings for organization, complexity, accuracy, and fluency. Second, because of the two misfit raters and the deletion of their evaluations, 11 out of 144 recordings were single-rated in the dataset, while the others were either double- or triple-rated. Third, the study adopted only a quantitative approach; hence, future studies may benefit from a mixed methodology that includes interviewing raters to better understand their perceptions of assessment.

Despite these limitations, the current findings provided important implications for evaluating opinion-based oral performances. First, as a research implication, the MFRM should be employed to assure quality and maintain high rater reliability (e.g., Aryadoust et al., 2021). In this study, nine raters assessed the participants' oral performances consistently, which was within the appropriate fit. This was a good way to examine how the criteria and the rubric can function reliably. Second, as a pedagogical implication, language teachers could take the constructs of non-linguistic features into consideration and possibly adapt the rubrics of this study for their classroom assessments. Among the linguistic features of CAF, fluency components such as length of speaking time or repair frequency were considered more influential when judging speaking performances from a rater's perspective. Therefore, fluency training can be promoted (e.g., Tran & Saito, 2021). Although this study does not deny its use of analytical CAF indices, it also hopes that other aspects such as the assessment of communicative achievement or speech organization, can be considered from a listener's point of view in the real world.

Conclusion

This study examined the human assessment of opinion-based oral performances and objective CAF measurements. Raters' perceived organization, complexity, accuracy, and

fluency were subjected to the MFRM. This means that this study's human ratings consisted of linguistic features (CAF) and a nonlinguistic feature (speech organization). Employing a slightly different assessment helped differentiate this study from previous ones, which used holistic ratings of overall L2 speakers' oral performances (e.g., Revesz et al., 2016; Suzuki & Kormos, 2020; Sato, 2012).

The FACET results provided some insights into understanding how the raters assessed the participants' oral performances. The regression analysis results showed that compared to complexity and accuracy, analytical measures of fluency strongly predicted human ratings, which was consistent with previous findings (e.g., Revesz et al., 2016; Suzuki & Kormos, 2020). Further studies must be conducted to reassess the use of the rubrics with different tasks and participant groups.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-022-00154-9.

Additional file 1. Appendix A Opinion Based Questions. Appendix B Human Rater Training.

Acknowledgements

I am grateful to the anonymous reviewers for their useful comments and to David Beglar and Kurtis McDonald for valuable comments on the analysis and statistics.

Author's contributions

Chie Ogawa is the solo contributor to this research paper. The author was responsible for the research design, data collection, data transcription, statistical analysis and approving the final manuscript.

Availability of data and materials

The datasets used during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 19 September 2021 Accepted: 6 January 2022
Published online: 14 February 2022

References

Arczis Web Technologies, Inc. (2019). *Syllable count*. http://www.syllablecount.com. Accessed Apr 2019.

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.

Boersma, P., & Weenink, D. (2009). PRAAT: Doing phonetics by computer [Computer Software]. Retrieved from http://www.praat.org. Accessed Mar 2017.

Bosker, H. R., Pinget, A., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394.

De Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533–568. https://doi.org/10.1111/j.1467-9922.2010.00620.x.

De Jong, N., Steinel, M., Florijn, A., Schoonen, R., & Hulstijn, J. (2012). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and nonnative speakers. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency investigating complexity, accuracy and fluency in SLA*, (pp. 121–142). Amsterdam, The Netherlands: Benjamins.

De Jong, N., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, *20*(3), 384–404. https://doi.org/10.1177/1362168815606161.

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS)*, (pp. 17–20). Stockholm: Royal Institute of Technology.

De Jong, N. H., Hulstijn, J. H., Schoonen, R., & Groenhout, R. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243. https://doi.org/10.1017/S0142716413000210.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*(5), 893–916. https://doi.org/10.1017/S0142716412000069.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*(4), 474–509. https://doi.org/10.1093/applin/amp042.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. Amsterdam, The Netherlands: Benjamins.
Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. https://doi.org/10.1093/applin/amm017.
Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information—Processing approach to task design. *Language Learning*, *51*(3), 401–436. https://doi.org/10.1111/0023-8333.00160.
Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*(3), 321–336. https://doi.org/10.1177/0265532216663991.
Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196. https://doi.org/10.1017/S0272263116000085.
Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.
Linacre, J. M. (2013). *Facets computer program for many-facet Rasch measurement, version 3.71.4*. Beaverton: Winsteps.com.
Linacre, J. M. (2017). *A user's guide to FACETS: Rasch-model computer programs (3.80.0) [computer software manual]*. Beaverton: Winsteps.com.
Magne, V., Suzuki, S., Suzukida, Y., Ilkan, M., Tran, M. N., & Saito, K. (2019). Exploring the dynamic nature of second language listeners' perceived fluency: A mixed-methods approach. *TESOL Quarterly*, *53*(4), 1139–1150.
McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392. https://doi.org/10.3758/BRM.42.2.381.
McDonald, K. (2018). Post hoc evaluation of analytic rating scales for improved functioning in the assessment of interactive L2 speaking ability. *Language Testing in Asia*, *8*(1), 19.
Nemoto, T., & Beglar, D. (2014). Developing Likert-scale questionnaires. In N. Sonda & A. Krause (Eds.), JALT2013 Conference Proceedings (pp. 1-8). Tokyo, Japan: JALT.
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.
Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CALF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044.
Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492.
Pallotti, G. (2009). CALF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045.
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079.
Revesz, A., Ekiert, M., & Torgersen, E. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828–848. https://doi.org/10.1093/applin/amu069.
Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, *355*(3), 866-900.
Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223–241. https://doi.org/10.1177/0265532211421162.
Sato, T., & McNamara, T. (2019). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics*, *40*(6), 894–916. https://doi.org/10.1093/applin/amy032.
Segalowitz, N. (2010). *The cognitive bases of second language fluency*, (1st ed.,). New York: Routledge. https://doi.org/10.4324/9780203851357.
Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38-62. https://doi.org/10.1093/applin/17.1.38.
Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93–120. https://doi.org/10.1111/1467-9922.00071.
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.
Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*, (5th ed.,). New York: Allyn and Bacon.
Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, *50*(2), 447–471. https://doi.org/10.1002/tesq.244.
Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task-performance in a second language*, (pp. 239–273). Amsterdam: Benjamins.
Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *TESOL Quarterly*, *50*(2), 369–393. https://doi.org/10.1002/tesq.232.
Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*(1), 84–119. https://doi.org/10.1093/applin/17.1.84.
Tran, M. N., & Saito, K. (2021). Effects of the 4/3/2 activity revisited: Extending Boers (2014) and Thai & Boers (2016). *Language Teaching Research* 1362168821994136.
Yan, X., Kim, H. R., & Kim, J. Y. (2021). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing*, *38*(4), 485–510. https://doi.org/10.1177/0265532220951508.

Publisher's Note