System 120 (2024) 103208

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

Predicting functional adequacy from complexity, accuracy, and fluency of second-language picture-prompted speaking

Rie Koizumi [a,*], Yo In'nami [b]

[a] University of Tsukuba, Japan
[b] Chuo University, Japan

ARTICLE INFO

Keywords:
Functional adequacy
Speaking ability
Syntactic complexity
Accuracy
Fluency
Japanese learners of English
Second-language speaking
Multiple regression

ABSTRACT

Functional adequacy (FA) is a construct of task achievement in communicative settings and focuses on the extent to which task requirements are satisfied by effectively conveying intended messages. Recent studies in second-language speaking and writing have emphasized the importance of FA in addition to complexity, accuracy, and fluency (CAF); FA and CAF are combined into the acronym CAFFA. This study aims to investigate (a) the extent to which CAF measures can explain FA holistic ratings in oral picture narration among Japanese learners of English; (b) how these results are moderated by different picture tasks; and (c) the comparability of means, variances, and correlations of the same FA ratings and CAF measures across tasks. Results of multiple regression analyses indicate that only a speed fluency measure (syllables per minute) significantly predicts FA, while a substantial proportion of FA remains unexplained by CAF, highlighting the separate and related nature of the two constructs. Moreover, the prediction of FA by CAF is consistent across picture tasks, with means and variances of FA and CAF measures being generally comparable, and with correlations of the same measures across tasks not being consistently strong, except for syllables per minute. The paper discusses implications and offers suggestions for future research.

1. Introduction

Second language (L2) speaking proficiency consists of multiple components. For example, De Jong (2023) proposed a model of L2 speaking proficiency based on previous studies (e.g., Bachman & Palmer, 2010), which comprised two components (linguistic and strategic), each of which includes knowledge and speed, and multiple subcomponents (e.g., structural, predictive, and pragmatic subcomponents of linguistic competence). Among a multicomponential and complex construct of L2 speaking proficiency, researchers have primarily assessed the linguistic aspects of complexity, accuracy, and fluency (CAF), typically using tasks that simulate real-life speech events (Housen, Kuiken, & Vedder, 2012). However, as Pallotti (2009) argued, such an exclusive focus on CAF could ignore essential parts of speech. Recent studies (e.g., De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012a; Kuiken & Vedder, 2017; Révész, Ekiert, & Torgersen, 2016) have expanded the focus to include the communicative aspects of how effectively intended messages are communicated, which is referred to as *communicative adequacy* or *functional adequacy* (FA), the latter of which is extensively used (e.g., Kuiken & Vedder, 2017; 2022a). Although assessing CAF and FA constructs together is a reasonable approach for comprehensively evaluating L2 speaking proficiency, the construct of FA remains underresearched; one promising area of investigation is to examine the

* Corresponding author.
E-mail address: koizumi.rie.ge@u.tsukuba.ac.jp (R. Koizumi).

https://doi.org/10.1016/j.system.2023.103208
Received 29 March 2023; Received in revised form 26 October 2023; Accepted 9 December 2023
Available online 26 December 2023

relationship between FA and CAF (Kuiken & Vedder, 2022a).

Therefore, this study aimed to investigate the extent to which CAF measures explain FA holistic ratings in oral picture narration among Japanese learners of English at novice to advanced levels.

2. Literature review

2.1. CAF and FA

In task-based assessment of L2 oral and written performance, CAF analysis using measures of quantifiable aspects has been the dominant approach since the 1990s (Kuiken & Vedder, 2022a), typically using ratio measures based on transcripts and countable aspects that machines or humans can judge fairly objectively (Housen et al., 2012; Koizumi, In'nami, & Jeon, 2022). Complexity refers to the degree to which L2 production incorporates an expansive and diverse array of advanced structures and lexicon (Housen et al., 2012). It comprises lexical, morphological, syntactic, and phonological complexity (Bulté & Housen, 2012). Accuracy is defined as the degree to which L2 production conforms to a norm (Housen et al., 2012). It comprises lexical, morphological, syntactic, and phonological accuracy. Fluency captures smooth flow in spoken language (Foster, 2020), comprising speed, breakdown, and repair fluency, that is, how fast L2 learners produce the L2, how they insert (or avoid) pauses, and how they insert (or avoid) correction, repetition, and false starts in L2 production (Révész et al., 2016; Suzuki & Kormos, 2020).

In addition to CAF, recent studies have highlighted the importance of FA in L2 speaking and writing assessment (e.g., De Jong et al., 2012a; Kuiken & Vedder, 2017; Révész et al., 2016). FA is defined as "a task-related construct, in terms of successful task completion by the speaker/writer in conveying a message to the listener/reader" (Kuiken & Vedder, 2022a, p. 1). FA is goal-oriented and comprises four dimensions: task requirements, content, comprehensibility, and coherence and cohesion. Task requirements relate to the extent to which task purposes and expectations are met successfully in terms of genre, task type, and other expected functions. Content refers to the extent to which ideas shown in learner production are acceptable and internally consistent. Comprehensibility specifies the extent to which listener/reader effort is needed to comprehend a speaker's/writer's intention and messages. Coherence and cohesion signify to what extent ideas and sentences in the production are adequately associated. Because high CAF features do not guarantee high FA, Kuiken and Vedder (2022c) argued that "assessing oral L2 performance is impossible without considering both CAF and FA (henceforth, CAFFA), and the mutual relationship and possible trade-offs between CAFFA dimensions" (p. 330). This is applied to assessing written L2 performance as well.

The heightened focus on FA, coupled with CAF, has garnered increasing interest within the realm of L2 acquisition. Notably, Pallotti (2009) stands as one of the pioneering scholars who advocated for an augmented emphasis on FA. A similar trend was observed in L2 assessment, in which the construct to be assessed expanded to incorporate the communicative aspect of ability and performance, such as "content" and "communicative effectiveness" (Sato, 2012). This shift in L2 assessment seems to trace back to McNamara (1996), who classified performance assessment into strong and weak types, with the strong type concentrating on the assessment of successful task fulfillment in real-world contexts, and the weak type concentrating on the assessment of linguistic aspects of performance.

To measure FA, rating scales that require human scoring were developed by Kuiken and Vedder (2017), Révész et al. (2016), and De Jong et al. (2012a). These rating scales are either holistic (De Jong et al., 2012a; Révész et al., 2016) or analytic, comprising the four dimensions mentioned above. These scales have been used in and adapted to various contexts (e.g., Pallotti, 2022; Strobl & Baten, 2022), with studies showing high interrater and intrarater reliability. For example, Kuiken and Vedder (2018) asked four nonexpert raters per task to use an FA analytic rating scale to evaluate texts from Dutch and Italian first-language (L1) and L2 speakers. They reported high interrater reliability. De Jong et al. (2012a) also reported high intrarater reliability across four nonexpert raters per task who used a holistic scale to evaluate oral texts from L1 and L2 speakers. The raters in these two studies underwent training to ensure that they were familiar with the FA construct, tasks, and rating scales, and to accurately score FA.

2.2. Relationships between FA and CAF

The relationships between FA and CAF can be interpreted as the effects of CAF on FA. While previous studies have typically used correlation or regression analyses, and results solely based on these methods do not establish causality, in the current case, it is plausible to consider CAF as affecting FA, rather than the other way around. This is because it is reasonably conceived that underlying linguistic knowledge and processing ability (e.g., knowledge of vocabulary and grammar; speed of lexical retrieval, articulation, and sentence building; and pronunciation; De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012b) affect the linguistic features of CAF, which are fairly independent from rater judgement; these CAF features, in turn, affect raters' perception toward the linguistic features, and raters consider the linguistic quality and evaluate the communicative adequacy of L2 speaking or writing. Although there may be a third set of confounding factors unrelated to L2 that affects both FA and CAF, which leads to spurious correlations, such factors are currently not conceivable.

Table 1 summarizes key studies on the relationships between FA and CAF, and the effects of task type and L2 proficiency on such relationships. Research on this topic is emerging, and so far, has provided mixed results (Kuiken & Vedder, 2022b). To the best of our knowledge, there are only five studies on this topic, with two focusing on speaking and three on writing; the current study focuses on speaking. First, Révész et al. (2016) analyzed L1 and L2 English speakers' speech during role-play tasks using a holistic FA scale and 32 CAF measures. Results based on linear mixed effects regression analyses showed that FA was predicted by 10 measures of complexity (both lexical and syntactic), accuracy, and fluency, with (breakdown and speed) fluency being the most significant dimension. This

Table 1
Relationships between FA and CAF in previous studies.

	Participants	Raters	Tasks	FA	CAF	Results
Révész et al. (2016), Speaking	100 English users; 20 L1 and 80 L2 speakers (40 L1 Japanese and 40 L1 Spanish learners; low-intermediate to advanced)	20 raters (10 with linguistics major and 10 with different majors; 2 raters evaluated each performance). With rater training	5 monologic role-play tasks (e.g., complaint and refusal; integrated tasks; computer-delivered	Holistic scale of 0–7	32 measures of SC, LC, A, and F	*Linear multilevel mixed effects regression analyses showed that 10 measures predicted FA significantly (R^2 = .41) e.g.,* F (breakdown F): Filled pause frequency: No. of filled pauses/100 words: R^2 = .15 (when analyzed separately) F (speed F): Mean duration of syllables: Speaking time (excluding pauses)/syllables = .07 A: General accuracy: No. of errors/100 words = .06 SC: Overall SC: No. of words/AS-unit = .06 LC: Lexical diversity: D = .05 F (breakdown F): Silent pause frequency: No. of silent pauses/100 words = .04 SC: subordination SC: No. of clauses/AS-unit = .04 *No interaction effect of task types* *Interaction effect of L2 proficiency* Fluency (repair fluency): No. of false starts/100 words. The effect was found only for advanced speakers.
Ekiert, Révész, Torgersen, and Moss (2022), Speaking	40 L1 Spanish learners of L2 English; low-intermediate to advanced)	Same as above	2 monologic role-play tasks (i.e., complaint and refusal; the rest was the same as above)	Same as above	4 fluency measures	*Linear multilevel mixed effects regression analyses showed one significant prediction* F (breakdown F): Silent end-clause pause frequency: Number of silent end-clause pauses/clause: R^2 = .237 *Interaction effect of task types* In the refusal task: F (breakdown F): Filled end-clause pause frequency: Number of filled end-clause pauses/clause = .09 *No interaction effect of L2 proficiency*
Kuiken, Vedder, and Gilabert (2010), Writing	103 L2 learners (L2 Dutch, Italian, and Spanish; L1 varied; mostly CEFR A2-B1)	4 Dutch raters, 3 Italian raters, and 3 Spanish raters	2 decision-making tasks at the CEFR B1 level	Holistic scale of 0–6	5 measures of SC, LC, and A	*Correlations with FA:* A: Errors/100 words = −.713 to −.473 A: Errors/T-unit = −.732 to −.199 LC: Guiraud index = .262 to .671 *Correlations between two tasks in Kuiken and Vedder (2017) using part of the data in Kuiken et al. (2010)*: Content = .607 to .623 Task requirements = .455 to .701 Comprehensibility = .766 to .877 Coherence and cohesion = .719 to .802
Pallotti (2022), Writing	217 writers: 64 L2 Italian (Various L1s; age 8–10); 153 L1 speakers of L2 Italian	10 nonexpert raters (Kuiken & Vedder, 2022b, p. 30)	1 narration task based on a 5-min video clip (i.e., tell the story to the teacher)	Analytic scale of 3 criteria from 1 to 6	2 measures of LC and F	*Correlations with FA:* F: No. of tokens = .43 to .61 LC: Lexical diversity; MATTR (moving average Type-Token Ratio) = .38 to 49

(*continued on next page*)

Table 1 (*continued*)

	Participants	Raters	Tasks	FA	CAF	Results
Strobl and Baten (2022), Writing	30 L1 Dutch learners of L2 German; B2 or C1 level	3 expert raters of (near-) L1 German; 2 raters evaluated each performance	2 personal narration tasks combined	Analytic scale of 3 criteria from 1 to 6	7 measures of SC, LC, A, and F	*Correlations with FA dimensions:* Content and topic development (CTD) & F (No. of tokens) = .675 CTD & LC (Guiraud index) = .616 Comprehensibility & A (No of error-free clauses/clause) = .311 Comprehensibility & LC (mean word length) = .289 Coherence & Cohesion (C&C) & F (No. of tokens) = .573 C&C & LC (Guiraud index) = .608 C&C & LC (CEFR-level-band value) = .277

Note. FA = Functional adequacy. CAF = Complexity, accuracy, and fluency. SC = Syntactic complexity. LC = Lexical complexity. F = Fluency. A = Accuracy. CEFR = Common European Framework of Reference for Languages.

means that all CAF aspects, especially fluency, were essential in explaining FA and that L1 and L2 speakers who pause more and speak more slowly tend to show lower FA.

Another study that explored the relationships between FA and CAF, and the effects of task types and L2 proficiency on such associations, was the work by Ekiert et al. (2022). They analyzed breakdown fluency further, using part of the data from Révész et al.'s (2016) study. This analysis employed linear multilevel mixed-effects regression techniques, with a focus on L2 learners of English who had an L1 Spanish background and performed two role-play tasks. They reported that the number of silent end-clause pauses per clause predicted FA substantially, suggesting that L2 learners who pause more at the end of the clause, without saying *uh* or *mm*, are likely to show lower FA. There was also an interaction effect of task types on the relationship between FA and the number of filled end-clause pauses per clause when performing refusal (but not when complaining). This indicated that L2 learners who insert more fillers at the end of clauses when they refuse tend to be perceived by raters as showing lower FA. The researchers interpreted these findings as indicating that refusal requires more sensitive speech, or "a fine-tuned and sequentially organized stretch of speech," and that raters may perceive superfluous end-clause fillers as less convincing and functionally inadequate (p. 52).

Three observations can be made based on the L2 production studies summarized in Table 1. First, FA was predicted by syntactic complexity to almost no degree, whereas fluency, accuracy, and lexical complexity seemed to be moderate predictors of FA. In particular, fluency seemed to be a key predictor, as reported, for example, by Révész et al. (2016) and Strobl and Baten (2022). However, results related to lexical complexity require further examination because a few such measures (e.g., Guiraud index and D) are known to be affected by text length, especially in the case of short texts, and may not accurately measure the intended construct of lexical diversity (Zenker & Kyle, 2021). Second, the relationship between FA and CAF may be moderated by tasks. Among studies that showed the comparison of learner performance across tasks, Ekiert et al. (2022) and Kuiken et al. (2010) reported conflicting results across tasks, whereas Révész et al. (2016) found consistent results. Although Ekiert et al. (2022) used part of the data from Révész et al. (2016), the results differed across the studies. This suggests that the relationship between FA and CAF seems to vary depending on several factors, including tasks (i.e., five vs. two role-play tasks), L2 learners' mother tongues (i.e., Japanese and Spanish vs. Spanish only), the range of language proficiency (i.e., whether studies include L1 speakers), and CAF measures (i.e., filled pause frequency vs. filled pause end-clause frequency). Third, these five previous studies have explored various languages, using multiple raters and both analytic and holistic FA scales, and employing various CAF measures. However, more studies, particularly in L2 speaking, are required to examine FA-CAF relationships in various contexts to better understand this issue. Accordingly, this study included L2 novice-to-advanced learners, using multiple tasks of simple picture narration.

2.3. Equivalency of picture prompts

Picture prompts are a valuable tool in L2 research and assessment for eliciting relatively long monologues while providing speech content (De Jong & Vercellotti, 2016). Rossiter, Derwing, and Jones (2008) considered picture stories an effective method "to maintain some control over the language elicited, while giving learners enough flexibility to provide us with a relatively realistic sample of their speaking proficiency" (p. 325). However, even with picture prompts that appear to have similar content, L2 learners tend to produce varied CAF features. For example, De Jong and Vercellotti (2016) compared the speeches of 25 high-intermediate learners of L2 English with various L1s, using five picture stories with six frames each, and similar content for sequential structures, storyline complexity, and main characters and props. They found substantial differences across picture tasks in fluency and lexical complexity, but not in accuracy and syntactic complexity. Inoue (2011) compared oral narratives of intermediate to advanced Japanese learners of L2 English (as per the Common European Framework of Reference for Languages [CEFR] B1–C1 levels, $N = 65$) using two picture tasks

with six frames each. She reported differences in accuracy and syntactic complexity, but not in fluency and lexical complexity. In a study with 20 Japanese learners of L2 English at novice to intermediate levels (CEFR A2–B2), Kakitani (2023) also examined differences across seven pictures with six frames each and found differences in syntactic and lexical complexity and accuracy measures, but not in fluency measures. The three previous studies reported differential effects of pictures on CAF, which may be explained by different features in each study, such as the L2 proficiency levels of learners and different parallel pictures. Two further points should be noted. First, although previous studies compared means between values from the same measures for different tasks, two additional conditions are required if researchers intend to argue that measures are strictly equivalent across similar tasks: (a) same standard deviations (or variances) between values from the same measures for different tasks, and (b) strong correlations between these values or the same correlations with other tests (Suzuki & Koizumi, 2021). Second, previous studies focused on CAF, and the comparability or equivalency across tasks in terms of FA have not been examined. This study examined the differences in means, variances, and correlations of FA and CAF features, while comparing picture prompts of relatively similar structures. Such an examination has not been conducted in the context of comparing picture prompts and is important because it addresses the question of the degree to which similar picture prompts are considered equivalent in terms of FA and CAF, and the extent to which results of a picture prompt can be applied to similar picture prompts (i.e., generalizability of the results of picture prompts; De Jong & Vercellotti, 2016).

2.4. Purposes and research questions (RQs)

This study examined how FA is predicted by CAF in the context of L2 narrative speaking among Japanese learners of English and how these relationships are moderated by different picture tasks. To assess the comparability of tasks, we also examined the means, variances, and correlations of each task for every measure. The study aimed to provide insights into dimensions of the speaking construct and how tasks affect FA and CAF.

We formulated the following research questions (RQs).

RQ1: To what extent does CAF predict FA in picture narration?
RQ2: To what extent does the prediction of FA by CAF differ across different picture tasks?
RQ3: To what extent do the means, variances, and correlations of FA and CAF differ across different picture tasks?

3. Method

3.1. Research design and participants

This study involved two types of participants: university students (n = 39) and senior high school (HS) students (n = 338; see Table 2). The former completed four tasks, whereas the latter completed only Task 2. The data of university students were used to examine all the RQs, whereas the data of HS students were analyzed separately to examine RQ1 more comprehensively.

The data analyzed were derived from a total of 377 Japanese learners of English, who were part of a larger project examining L2 learners' speaking development longitudinally and cross-sectionally (see Koizumi & In'nami, 2022)[1]. The students were from two universities (n = 39) and two public HSs (n = 338) and had learned English as a foreign language at Japanese secondary and tertiary schools for 3–12 years. Their exposure to the L2 was primarily limited to classroom instruction and self-learning settings. The L2 English proficiency levels of the university and HS groups were approximately at the CEFR A1–C1 and A1–B1 levels, respectively, based on their learning experiences, teachers' judgements, and English proficiency tests that the teachers reported they had passed (Eiken Test Grades 1–3; see https://www.eiken.or.jp/eiken/en/research/comparison-table.html). The participant details are provided in Appendix Table A1.

3.2. Speaking tasks and test procedures

The speaking test was intended to assess and diagnose L2 speaking ability for low-stakes classroom assessment or research purposes. The test comprised four tasks of picture narration (or description; Tasks 1–4). Each task (i.e., picture prompt) comprised a sequence of two frames taken from interview tasks in the Eiken Test Grade 2, with an estimated difficulty of CEFR B1, developed and administered by the Eiken Foundation of Japan (https://www.eiken.or.jp/eiken/en/research/comparison-table.html)[2]. These tasks

Table 2
Relationships between participant, tasks, and research questions.

	Task 1	Task 2	Task 3	Task 4
University students (n = 39)	✓	✓	✓	✓ RQ2 & RQ3
High school students (n = 338)		✓ RQ1		

were selected based on the L2 proficiency levels of HS students and were relatively easy for some university students[3]. HS students performed Task 2, whereas university students performed Tasks 1–4, with the order counterbalanced to avoid the order effect.

Although the difficulty and structure of the four tasks were generally considered comparable, there were some differences, especially in terms of sequential structures and storyline complexity. Tasks 1 and 2 showed a predictable story with smooth transitions. Task 3 showed a predicament in the second frame (i.e., character spending all their money on bargain-priced sports items, with no money left to take the bus), and Task 4 had a predicament in the first frame (i.e., character having no flowers to give to their wife, as the flower shop they had in mind was closed, and then finding a flower vending machine that operates 24 h a day). These differences were difficult to control because comparisons between picture prompts were conducted on an ad hoc basis.

The speaking test was conducted in person in a silent room by an interviewer who was either a schoolteacher, researcher, or research assistant. Each examinee was given 1 min to plan their speech and was asked to describe the frames in one or four tasks for 2–3 min. Most examinees finished their narration within 1 min. The interviewer recorded the speech using either a tape or digital recorder.

3.3. FA rating scale

The FA rating scale was originally based on De Jong et al.'s (2012a) study and subsequently modified by Koizumi and In'nami (2022). As seen in Tables 3 and 4, the FA scale[4] ranged from levels 1–4 and included four task-specific content points, which were necessary to meet the task requirements for successful picture narration and to ensure story coherency. The scale included the dimensions of content, comprehensibility, task requirements, and coherence and cohesion, in line with Kuiken and Vedder's (2017) study. Specifically, the former two received greater attention.

3.4. CAF measures

Table 5 shows the five CAF measures that were calculated. They were selected based on previous studies (e.g., Koizumi et al., 2022; Révész et al., 2016; Suzuki & Kormos, 2020) and correlational results with similar measures (see Appendix Figure A1 and Table A2). Fluency was assessed using two measures: syllables per minute (speed fluency) and disfluency markers per 100 words (repair fluency). Breakdown fluency, as measured by the number or length of pauses, was not included because of poor recording conditions. Syllables per minute was selected because it is a more precise measure of speed fluency than the number of words per minute, and because the two measures were strongly correlated ($r = 0.98$). Disfluency markers included verbatim repetitions, self-corrections, and false starts, based on Foster et al.'s (2000) study. "Words" meant pruned tokens, which were counted after excluding disfluency markers and filled words such as *mm* and *ah*. Disfluency markers per 100 words was selected because it showed a stronger correlation with FA ($r = -0.28$) than the two other repair fluency measures (disfluency markers per clause, $r = -0.26$; disfluency markers per minute, $r = 0.04$), and because the three measures showed strong correlations with each other ($r = 0.72$–0.89).

Accuracy was measured using error-free clauses per minute. The number of errors was not counted because numerous errors were difficult to identify as discrete units.

Complexity (syntactic complexity) was measured using AS-unit (Analysis of Speech Unit) length, and clauses per AS-unit. An AS-unit is defined as "a single speaker's utterance consisting of *an independent clause, or subclausal unit*, together with any *subordinate clause* (s) associated with either" (Foster, Tonkyn, & Wigglesworth, 2000, p. 365). Clause length was excluded because it correlated with FA weakly ($r = 0.01$), correlated only moderately with AS-unit length and clauses per AS-unit ($r = 0.58$ and -0.36, respectively), and caused severe multicollinearity in the regression (variance inflation factor [VIF] = 13.27–46.99).

Lexical complexity was not measured because the speech was too short to compute the same. For example, in Task 2 ($N = 377$), 74.01% of the participants produced texts of less than 50 words, with a minimum of 5 words. Zenker and Kyle (2021) reported that 50 words was the minimum required to use the moving average type-token ratio (MATTR).

Table 3
FA scale.

Level	Overall description	Descriptors	Content points
4	A successful contribution	The speaker clearly communicates essential elements. The response meets the minimum requirements. **And** the response is coherent. It can be easily understood.	4 items
3	A moderately successful contribution	The speaker communicates essential elements to a limited extent. **And/or** the response is somewhat coherent. It is possible to understand the response with some effort.	3 items
2	A weak contribution	The speaker communicates essential elements to a very limited extent. **And/or** the response lacks coherence, and it is difficult to understand.	2 items
1	An unsuccessful contribution	The speaker does not communicate or hardly communicates essential elements. **And/or** the response lacks coherence, and it is very difficult to understand.	1 item

Note. Levels can be raised upward when there is a notable indication of much better speech (e.g., the provision of detailed information) even when certain content points are not mentioned.

Table 4
Content points in each task used in FA scale.

Task	Content points (four items for each)
1: Volunteer activity	(a) A woman finds an article on a volunteer activity. (b) There are a man and a woman. (c) They join the volunteer activity. (d) People are cleaning.
2: Birthday present	(a) The grandmother's birthday is approaching. (b) Two grandchildren select a present for her. (c) The grandmother receives a bag. (d) She and her grandchildren are talking.
3: Outlet shopping	(a) A man buys some sports items. (b) There is a sale. (c) He uses all his money. (d) He cannot take a bus.
4: Vending machine	(a) The flower shop is closed. (b) It is the man's wife's birthday. (c) He finds a machine. (d) He buys flowers for his wife.

Table 5
CAF measures.

Construct	Label	Definition
Fluency	F1SPM: Syllables per minute (Speech rate; speed fluency)	The number of syllables divided by the total speaking time (including pauses), multiplied by 60
	F2DPW[a]: Disfluency markers per 100 words (Repair fluency)	The number of disfluency markers divided by the number of words, multiplied by 100
Accuracy	A: Error-free clauses per clause	The number of error-free clauses divided by the number of clauses
(Syntactic) complexity	SC1WAS: AS-unit length (Overall SC)	The number of words divided by the number of AS-units
	SC2CAS: Clauses per AS-unit (Subordination SC)	The number of clauses divided by the number of AS-units

[a] Lower values mean more desirable features.

3.5. Scoring and coding

Three L1 Japanese raters—the two authors and a research assistant (Raters 1–3)—scored FA. All had extensive experience in teaching and assessing L2 speaking. They listened (without transcripts) to and evaluated the first minute of speech for each task using the FA rating scale. All responses were scored independently and separately for each task by two or all three of the raters. They underwent a 3-h rater training; as is typically conducted in such rater training (e.g., Knoch, Fairbairn, & Jin, 2021), the trainees carefully read the scoring guidelines, scoring sample performances by themselves and discussing and modifying the scoring processes and outcomes to understand the construct of FA and how it is operationalized in the rating scale.

Before computing CAF measures, the first minute of speech for each task was transcribed because 1 min was sufficient for most participants to complete the speech[5]. The first author, along with two research assistants majoring in English, first transcribed the speech and after an interval of one month or more, the first author carefully checked and corrected the first transcription.

Two researchers (the first author and another coder, who were both L1 Japanese speakers) underwent a training session (similar to the one for FA scoring) and coded 10% of the whole speeches in terms of aspects that required coders' judgments (e.g., the number of clauses with errors). The researchers demonstrated high intercoder reliability ($r = 0.91$–1.00), and resolved discrepancies in scoring through discussion. The remaining speeches were coded by the first author, who counted necessary elements (e.g., the number of syllables), using the KWIC (Key Word in Context) Concordance for Windows (http://nuchs-corpus.japanwest.cloudapp.azure.com/kwic/) and Text Inspector (https://textinspector.com/), and calculated the CAF measures using Microsoft Excel. After approximately one month or later, the first author double checked the coding and calculation.

3.6. Analyses

The FA raw scores from the four tasks were analyzed using many-facet Rasch measurement. Koizumi and In'nami (2022) reported high reliability with regard to the participants and tasks (0.72 and 0.91, respectively) and adequate model fit for the data using FACETS (Version 3.83.6; https://www.winsteps.com/facets.htm). Rasch-estimated logit scores (termed FA Rasch scores) were used in this study, because they reflected FA as a construct independent from tasks and they showed a very strong correlation with average human-rated raw scores for each task (e.g., $r = 0.98$; see Appendix Figure A1).

To answer RQ1 and RQ2, simultaneous (forced-entry) multiple regression analysis was conducted with the learners' FA Rasch scores as a dependent variable and with CAF values as independent variables.

Regarding the statistical assumptions for regression analysis (Field, Miles, & Field, 2012), multicollinearity was not found, with VIF values ranging from 1.15 to 2.91, which were below the threshold of 10. Four participants were identified as outliers, with a standardized residual of −2.88 to 2.67. We did not exclude such outliers because the robust regression analysis, which excluded outlying data, showed very similar results. All analyses were conducted using R statistical software (Version 4.1.2) and were based on Field et al.'s (2012) textbook. Supplementary materials are available via the Open Science Framework (https://osf.io/kxw35/).

We conducted power analysis and calculated the required sample size using G*Power (Version 3.1.9.7; https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower; Faul, Erdfelder, Buchner, & Lang, 2009). The sample size required was 43, in case of testing F tests, multiple regression: fixed model, R^2 deviation from zero with large effect size, $f^2 =$

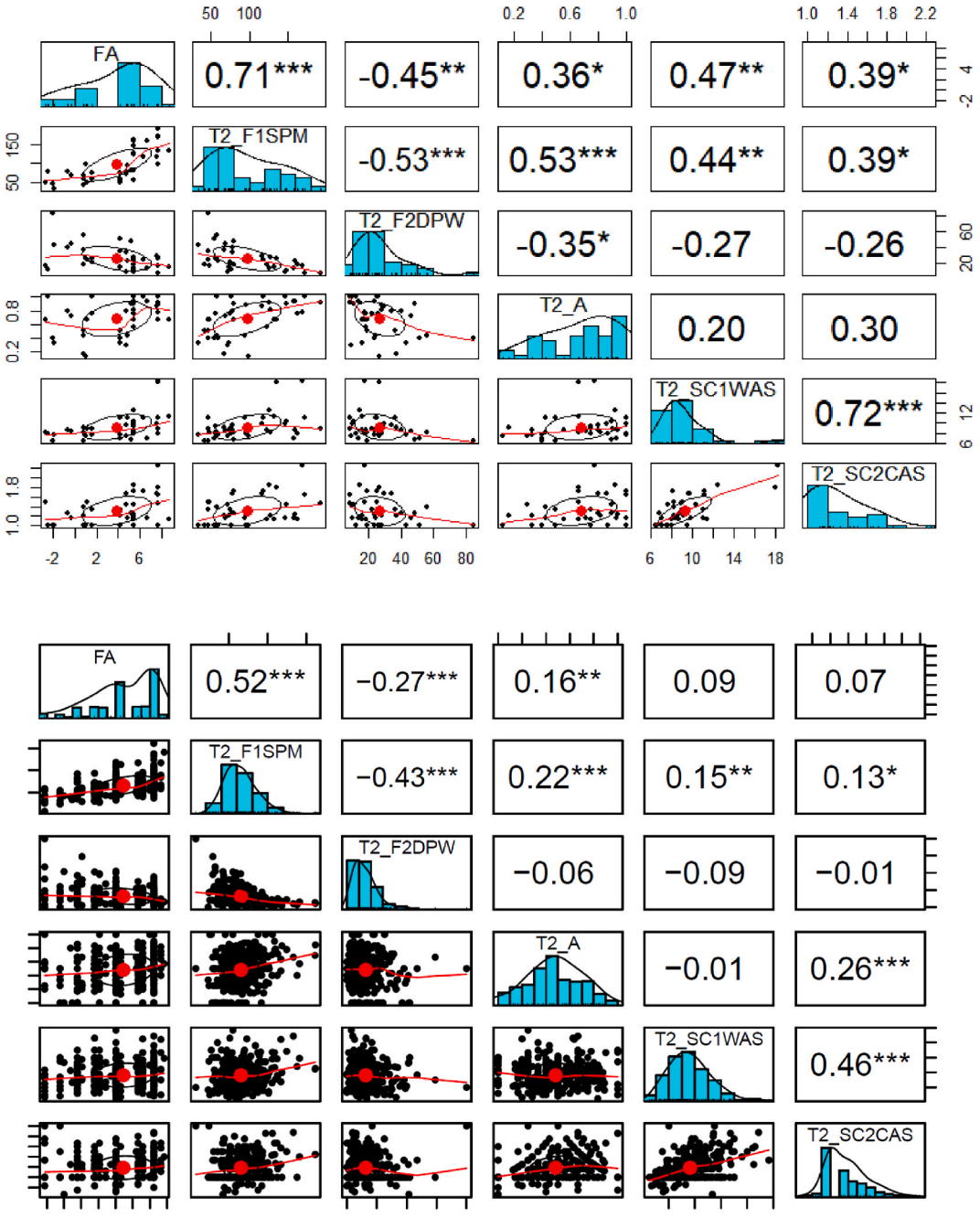

Fig. 1. Correlations among the variables of the university and high school groups (n = 39, top; n = 338, bottom; Task 2).
Note. See Table 6 for the variable names.

0.35 (i.e., $\rho^2 = 0.26$), $\alpha = 0.05$, $1 - \beta$ (power) = 0.80, and the number of predictors = 5 (see Field et al., 2012). For the current analyses, the actual sample sizes of 39 and 338 for university and HS students, respectively, close to the required sample size or larger, were considered acceptable.

To examine RQ3, we used repeated measures analysis of variance (ANOVA) and Pearson product-moment correlation coefficients to investigate the degree of differences in the means, variances, and correlations of FA ratings and CAF measures.

4. Results

Table 6 shows the means and standard deviations for each variable (see Appendix Table A3). Fig. 1 shows the relationships between

Table 6
Means (and standard deviations) for each variable from 39 university students and 338 high school students (HSS), and the whole participants (N = 377).

	University students				HSS	Whole
Task	1	2	3	4	2	2
FA Rasch scores[b]		3.54 (3.71)			2.93 (3.57)	3.02 (3.54)
FA raw scores	3.49 (0.74)	3.27 (0.83)	3.37 (0.70)	3.53 (0.62)	3.17 (0.80)	3.18 (0.80)
F1SPM: Syllables per minute	105.06 (44.98)	96.86 (44.01)	97.81 (44.86)	102.11 (45.94)	65.37 (22.28)	68.63 (27.08)
F2DPW[a]: Disfluency markers per 100 words	21.33 (15.81)	26.91 (15.34)	22.59 (17.71)	26.55 (18.97)	31.38 (23.72)	30.92 (23.02)
A: Error-free clauses per clause	0.65 (0.27)	0.68 (0.25)	0.67 (0.27)	0.63 (0.26)	0.48 (0.23)	0.50 (0.24)
SC1WAS: AS-unit length	8.65 (1.95)	9.25 (2.54)	9.44 (2.43)	8.50 (1.94)	7.63 (1.61)	7.80 (1.79)
SC2CAS: Clauses per AS-unit	1.22 (0.30)	1.31 (0.30)	1.42 (0.33)	1.35 (0.23)	1.19 (0.23)	1.20 (0.24)

Note.
[a] Lower values mean more desirable features.
[b] Computed using FA ratings of Tasks 1 to 4.

the variables based on the university (n = 39) and HS (n = 338) students' data (see also Appendix Figures A.2–A.4).

As summarized in Table 7, the regression analysis with 39 university students using five CAF measures (Model 1) was found to be useful in predicting FA, F (5, 34) = 8.58, p < .001. Specifically, 55.78% (R^2 = 0.5578) of the variance in FA was accounted for, with syllables per minute (a speed fluency measure) contributing significantly to the prediction (β = 0.32). When only syllables per minute was used for prediction (Model 2; β = 0.41), the results remained significant, explaining 51.12% of the variance in FA (see Appendix Table A4 for results of Tasks 1, 3, and 4). These results indicate that the remaining four variables contributed to a very limited degree (4.66% [55.78% minus 51.12%]); a large percentage of the variance in FA was not explained by the five CAF measures (44.22% [100% minus 55.78%]), and therefore, remained unexplored.

In the case of the HS data, the results were similar (see Table 8). Syllables per minute was the sole and most effective predictor of FA, explaining 28.05% and 27.46% of the variance in FA (β = 0.49 and 0.52, for Models 1 and 2, respectively). The remaining four CAF measures further explained 0.59% (28.05% minus 27.46%) of the variance in FA.

Table 9 summarizes the percentages of the variances in FA that were predicted by CAF. In Task 2, FA in the university data was predicted by CAF measures (55.78%) more than in the HS data (28.05%). Across Tasks 1–4, the patterns observed in university students were consistent, with syllables per minute solely predicting FA substantially (51.12–58.55%) and with a large percentage of the variances unexplained by CAF (37.67–44.22%). Thus, the results consistently suggested a moderate tendency in which those who produced more words and syllables per minute and spoke more fluently likely obtained higher FA scores, probably because they could produce more thorough descriptions of the picture items.

In Task 2, the percentages of variance in FA explained by the five CAF measures (55.78%) and by syllables per minute (51.12%) among the university students were larger than those observed with the HS students (28.05% and 27.46%, respectively). The results may reflect differences in the target learners (more lower-level learners with a narrow range of proficiency, from the data of 338 participants).

Repeated measures ANOVA was used to examine whether the means of university students for each measure (i.e., five CAF measures plus FA raw scores) varied across Tasks 1–4 (RQ3; n = 39; see Appendix Table A5 for detailed results). Before the analysis, the statistical assumption was examined using Mauchly's test for sphericity for each measure, which indicated that the assumption of sphericity had been violated for clauses per AS-unit, w = 73, p = .046; thus, the variances of the differences between the tasks were not equal for clauses per AS-unit. Nonetheless, the assumption of sphericity had been met for the other five measures. Repeated measures ANOVA with a Greenhouse-Geisser correction showed that means of clauses per AS-unit differed across tasks, F (2.61, 99.18) = 4.40, p < .01, generalized η^2 = 0.05. Post hoc tests using the Bonferroni correction showed that clauses per AS-unit were smaller in Task 1 than in Tasks 3 and 4. In contrast, no significant differences were observed across tasks in the means and variances of the other five measures. Correlations of the same measures across tasks among university students were strong for syllables per minute (r =

Table 7
Regression results for FA of university students (Task 2, n = 39).

Model	Variable	B	95% CI of B	Standard Error of B	β
1	(Intercept)	−1.68	−6.45, 3.08	2.34	−.26
	F1SPM	0.04***	0.02, 0.07	0.01	.32
	F2DPW	−0.02	−0.08, 0.03	0.03	−.16
	A	0.16	−3.53, 3.85	1.82	.01
	SC1WAS	0.32	−0.12, 0.76	0.21	.16
	SC2CAS	−0.76	−4.27, 2.76	1.73	−.05
2	(Intercept)	−1.46	−3.28, 0.36	0.90	−.22
	F1SPM	0.05***	0.04, 0.07	0.01	.41

Note. CI = confidence interval. R^2 (Adjusted R^2) = 0.5578 (0.4928) and 0.5112 (0.4984) for Models 1 and 2, respectively. Model 1: F(5, 34) = 8.58, p < .001. Model 2: F(1, 38) = 39.74, p < .001.
*p < .05. ***p < .001.

Table 8
Regression results for FA of high school students (Task 2, $n = 338$).

Model	Variable	B	95% CI of B	Standard Error of B	β
1	(Intercept)	−2.39*	−4.64, −0.15	1.14	<.001
	F1SPM	0.08***	0.06, 0.09	0.01	.49
	F2DPW	−0.01	−0.02, 0.01	0.01	−.06
	A	0.90	−0.62, 2.42	0.77	.06
	SC1WAS	0.06	−0.17, 0.30	0.12	.03
	SC2CAS	−0.35	−2.03, 1.34	0.86	−.02
2	(Intercept)	−2.56***	−3.57, −1.55	0.51	<.001
	F1SPM	0.08***	0.07, 0.10	0.01	.52

Note. R^2 (Adjusted R^2) = 0.2805 (0.2696) and 0.2746 (0.2725) for Models 1 and 2, respectively. Model 1: $F(5, 332) = 25.88$, $p < .001$. Model 2: $F(1, 336) = 127.2$, $p < .001$.

Table 9
Percentages of FA predicted by CAF (R^2 x 100).

	University students				HSS
Independent variables	Task 1	Task 2	Task 3	Task 4	Task 2
Five CAF measures (A)	62.33	55.78	59.64	57.51	28.05
F1SPM: Syllables per minute rate only (B)	58.55	51.12	55.34	53.42	27.46
Four other CAF measures (A – B)	3.78	4.66	4.30	4.09	0.59
Remaining (1 – A)	37.67	44.22	40.36	42.49	71.95

Note. HSS = High school students.

0.75–0.90); however, for the other five measures, the correlations were small to moderate (e.g., $r = 0.29$–0.70 for FA raw scores; see Appendix Table A6).

5. Discussion

5.1. To what extent does CAF predict FA in picture narration? (RQ1)

The results showed that CAF predicted FA to a substantial degree, with more percentages being predicted by university students than by HS students. The primary predictor among the CAF measures for both groups was speed fluency (i.e., speech rate, or syllables per minute). However, a considerable percentage of the variance in FA remained unexplained by the CAF measures, suggesting that FA is a construct that measures relatively different aspects of L2 speaking ability from CAF (see Table 9).

The results that fluency substantially explained FA and that syntactic complexity was not a primary predictor aligns with previous studies on FA (see Table 1), suggesting that there is a moderate tendency in which those who speak more fluently tend to produce functionally adequate speech. This tendency has also been observed in L2 speaking studies on comprehensibility (or listeners' ease of understanding speech, which is an essential component of FA). In these studies, fluency was shown to be a contributing factor to comprehensibility (Saito, Trofimovich, & Isaacs, 2017). However, previous studies on L2 speaking and FA (Ekiert et al., 2022; Révész et al., 2016) showed that breakdown fluency, which is related to pauses, was the dominant fluency dimension, and speed fluency only played a minor role. By contrast, this study found that speed fluency was dominant. These differences may be related to three main factors: First, the task type used in previous studies on L2 speaking and FA was role play that required the accomplishment of pragmatic functions, whereas this study used a picture narration task type, which was simpler. Additionally, speaking more in picture narration likely led to better accomplishment of conveying the content and obtaining better FA. This first explanation seems to be supported by studies on L2 writing and FA (Pallotti, 2022; Strobl & Baten, 2022), which have used a simple task type, such as video-based and personal narration, and found moderate correlations between FA and the number of tokens (similar to a speed fluency measure; see Table 1). Crowther, Trofimovich, Saito, and Isaacs (2018) also reported the effects of task types on the relationships between comprehensibility and speech rate. The results of this study and those of previous studies suggest that the choice of task type may influence the relationship between FA and CAF.

Second, L2 proficiency may have played a stronger role in the correlations between FA and speed fluency in this study than in previous studies on L2 speaking and FA (Ekiert et al., 2022; Révész et al., 2016), which targeted L2 learners at low-intermediate to advanced levels. They did not include novice learners and had a narrower range of proficiency than the participants in our study (novice to advanced levels). Learners with lower proficiency may struggle to convey their message, and using more words speedily may have affected FA more. A similar finding was observed by Saito, Trofimovich, and Isaacs (2016), who reported that speech rate differentiated learners with lower-level comprehensibility better than it differentiated those with higher-level comprehensibility. Similarly, Pallotti's (2022) L2 writer participants (aged 8–10 years) had presumably low proficiency, and the number of tokens (similar to a speed fluency measure) moderately correlated with FA. Along with the inclusion of learners with lower proficiency, a wider range of proficiency levels may have impacted the results because in this study, the percentage of FA predicted by CAF was larger in the university group than in the HS group. Both groups included novice learners but differed in the proficiency range—The university group

Table 10
Example of narration in Task 2.

Rika and … Risa {eh} sinking about … grandmother's birthday present in two weeks later. / … Two weeks laters, {eh, Ri} Risa and … Rika {eh} received … {ah s} birthday present received at {grandfather ah} grandmother. / … Grandmother … {ta talking … ah} talking … telephone … to {eh} Rika and Risa / {um … Grandmo} grandmother … is very happy.

Note. 49 s { } = disfluency marker. His FA ratings and CAF measures in Task 2 were as follows: FA = 2.5 out of 4 (FA Rasch score = −0.66); syllables per minute = 78.37; disfluency markers per 100 words = 40.00; error-free clauses per clause = 0.80; words per AS-unit = 8.75; clauses per AS-unit = 1.25.

covered a wider range from CEFR A1 to C1 than the HS group from A1 to B1. Thus, the results suggest that the degree of FA explained by CAF measures may vary depending on the range of proficiency levels with and without novice learners.

Third, we were not able to include breakdown fluency measures for technical reasons; including them may have predicted FA more than speed fluency measures and the overall prediction may have been better. Task types, proficiency levels, and measures used may also explain differences between this study and previous FA studies in terms of FA's relationships with accuracy or lexical complexity.

Additionally, although speed fluency predicted a substantial portion of the variance in FA, CAF measures failed to explain a considerable portion. To show how linguistic features (other than speed fluency) correlate with FA, an example is presented from an undergraduate student, whose FA was insufficiently explained by syllables per minute, in Table 10.

In the Task 2 prompt, a grandmother received a birthday present from her grandchildren, Risa and Rika; however, the description provided by the speaker was the opposite (*Risa and Rika received birthday present*; this was a major syntactic error). Although the student tried to correct their utterance later, they were unsuccessful. Additionally, the speaker mispronounced the keyword *thinking* as *sinking*, causing a major pronunciation error and leading to difficulty in conveying the intended meaning. Furthermore, the speaker frequently paused and self-corrected, making the speech more difficult to comprehend. Although the speaker tried to describe the picture in more detail and at a faster pace, they obtained a lower FA rating. It should be noted that a collective occurrence of errors, disfluency markers, and other elements, each of which belong to a different category used for a different measure (e.g., error-free clauses per clauses; disfluency markers per 100 words), can contribute to major incomprehensibility in an integrated manner, even if the contribution of each element is not high. For example, in the abovementioned case, errors and frequent repetitions and pauses required intense concentration and effort, resulting in lower comprehensibility. This finding aligns with L2 comprehensibility studies, which suggest that comprehensibility is affected by various linguistic features of L2 speech, such as pronunciation, fluency, vocabulary, grammar, and discourse (Saito et al., 2017). FA, a broader and more multifaceted concept than comprehensibility, is also impacted by various elements that may not always be well captured by CAF measures.

5.2. To what extent does the prediction of FA by CAF differ across different picture tasks? (RQ2)

The relationships between FA and CAF were found to be consistent across the four tasks that had relatively similar structures (see Table 9). Despite small variations across the tasks, there were similar percentages of FA explained by CAF (55.78–62.33%) and speech rate (51.12–58.55%). This suggests that FA and CAF are related but different constructs that can be constantly measured across different picture tasks. In previous studies, significant effects of tasks were found in Ekiert et al.'s (2022) and Kuiken et al.'s (2010) studies, whereas nonsignificant effects were found in Révész et al.'s (2016) study (see Table 1). All three studies used relatively comparable tasks (e.g., decision-making tasks in Kuiken et al.'s [2010] study). The presence and absence of task effects seems challenging to predict; however, in terms of picture narration, which can control the content and output relatively well, it is likely that comparable results can be derived.

5.3. To what extent do the means, variances, and correlations of FA and CAF differ across different picture tasks? (RQ3)

Although the relationships between FA and CAF were similar across picture tasks, more rigorous analyses of measures can help practitioners understand the degree of task comparability for FA and CAF measures. A summary in Table 11 shows that the means and variances of clauses per AS-unit were different in this study whereas those of the other five measures were similar; the correlations between the values of the same measures using different tasks were strong for syllables per minute, while the correlations were weak to moderate for the other five measures. Because the four tasks used in the current study differ in terms of sequential structures and storyline complexity, obtaining comparable results in terms of means and variances in FA raw scores and four CAF measures may be considered positive. Regarding the varied means and variances of clauses per AS-unit, Tasks 3 and 4 produced syntactically more complex speech than Task 1, possibly because the former two depicted an accident and may have pushed learners to explain relatively complex circumstances using more clauses (Robinson, 2005), which led to varied performances. Regarding the correlations, strong ones were found in syllables per minute. Kuiken and Vedder (2017) presented strong correlations of FA analytic ratings between two decision-making tasks ($r = 0.719$–0.877 in comprehensibility, and coherence and cohesion; see Table 1). This contrasted with the weak to moderate FA correlations across tasks ($r = 0.29$–0.70), suggesting that task types may affect comparability of FA ratings and CAF

Table 11
Examining comparability across the picture tasks ($n = 39$).

	Means (this study)	Means (previous studies)	Variances (this study)	Correlations (*r*: this study)
FA raw scores	Similar	–	Similar	.29–.70
F1SPM: Syllables per minute	Similar	I: Similar K: Similar	Similar	.75–.90
F2DPW: Disfluency markers per 100 words	Similar	–	Similar	.32–.64
A: Error-free clauses per clause	Similar	D&V: Similar I: Different	Similar	.46–.74
SC1WAS: AS-unit length	Similar	D&V: Similar I: Similar K: Different	Similar	.19–.55
SC2CAS: Clauses per AS-unit	Different (T1 < T3 & T4)	D&V: Similar K: Similar	Different	.29–.58

Note. – = not examined. D&V = De Jong and Vercellotti (2016); I = Inoue (2011); K = Kakitani (2023). Similar = not statistically different across tasks. T = Task.

measures across similar tasks.

Syllables per minute showed similarities for the means, variances, and correlations across tasks, suggesting that very stable comparability could be expected when similar picture prompts are used. For the other five measures, caution may need to be exercised to consider equivalence across picture prompts.

Previous studies (De Jong & Vercellotti, 2016; Inoue, 2011; Kakitani, 2023) have compared means of the same CAF measures, and we used some of these measures in this paper (see Table 11). The results suggest similarities and differences across studies; it seems that subtle differences may affect task equivalency, and creating strictly equivalent tasks may be challenging. However, there was one exception: Syllables per minute (speech rate) consistently provided the same results, indicating that speech rate is likely to be comparable across picture prompts. This comparability may be related to strong correlations between speech rate and L2 speaking proficiency in general (Koizumi et al., 2022).

6. Conclusion

This study examined the degree to which CAF can predict FA in picture narration among Japanese novice-to-advanced learners of English (RQ1). The results showed that speech rate (syllables per minute) is the only significant predictor of FA and that a large proportion of FA is unexplained by CAF, indicating that these two are related but separate constructs. We also investigated the comparability of FA and CAF results and found that the prediction patterns of FA by CAF were consistent across picture tasks (RQ2). We further found that syllables per minute satisfied the condition of task equivalency, with similar means and variances and strong correlations of the values across tasks (RQ3).

The study's findings have important implications for practice, highlighting the need to assess FA in addition to CAF, because of FA's limited predictability through CAF measures. Additionally, it is necessary to raise teachers' and learners' awareness regarding the importance of efficacy of message conveyance in communication. In particular, teachers may need to receive training, and consider incorporating instructions related to FA and fluency (the most significant predictor of FA) and including FA for scoring rubrics (Ekiert et al., 2022). Furthermore, effectively measuring FA might be challenging for usual automated scoring systems because such systems may not capture FA's deeper meaning and functional aspects of communication beyond surface text information (Isaacs, 2018).

Although this study provides insights into constructs of FA and CAF and shows the consistency of results across picture prompts, it has three main limitations that should be addressed in future research. First, we only used one task type (picture narration), five key CAF measures, and one FA holistic rating scale. It is necessary to include (a) different task types with adequate difficulty that elicit various speaking aspects (e.g., interaction) and pragmatic and various other functions; (b) various CAF measures of, for example, breakdown fluency, lexical complexity, and weighted accuracy (Foster & Wigglesworth, 2016); and (c) different types of FA scales, to cover the wider construct of CAF and FA. These assessments may include human ratings and/or use tools that allow for precise measurement such as Praat (Boersma & Weenink, 2023). Second, we collected the data using a convenience sampling approach, and the current results may not reflect the whole population of Japanese high school and university students. Third, the university data were not very large, the HS students did not work on all tasks, and the speech elicited was short with low recording quality for detailed analysis. Future studies should explore a research design using a more representative group of learners who perform multiple tasks and longer speech samples with good recording quality that would allow more comprehensive analyses, for an in-depth understanding of L2 speaking proficiency.

Note

1. Previous studies in the same project (see Koizumi & In'nami, 2022) did not address the relationships between FA and CAF.
2. The Eiken Test Grade 2 has a written-test phase and an oral-interview phase that uses a picture prompt. In 2004, teachers at one HS chose four picture prompts, which were officially administered in 1998 and 2000. It should be noted that the current version of a picture prompt has a sequence of three, not two, frames with a tight structure (see https://www.eiken.or.jp/eiken/exam/grade_2/).

3. Although tasks were relatively easy for some university students and the FA rating scale had four levels, there were no signs of ceiling effects in FA (as well as CAF; see Appendix Table A3), partially because FA raw scores were scaled into FA Rasch scores using many-facet Rasch measurement, which allowed for a wide range of precise measurement.
4. The FA scale initially had eight levels, which was reduced to four based on the results of many-facet Rasch measurement, which showed a limited use of Levels 5–8.
5. In previous analyses (Koizumi & In'nami, 2022), a small number of university students ($n = 5$) talked about the first picture frame in detail and did not finish the narration within 1 min. Such students were not included, so there were no students who received a lower FA rating due to the detailed description without mentioning content points.

Funding

This work was funded by JSPS KAKENHI Grant Numbers 22720216, 20K00894, and 23K00753 and Tokiwa University Grant for Research Projects (2007–2008).

Author contributions

Rie Koizumi: Conceptualization, Funding acquisition, Methodology, Data collection, Data analysis, Writing (Original draft preparation). Yo In'nami: Methodology, Data analysis, Visualization, Writing (Reviewing and Editing), Funding acquisition.

Data availability statement

Although the original data cannot be shared due to ethical reasons, the simulated data created from the original data that allows researchers to reproduce the findings will be available from Open Science Framework (https://osf.io/kxw35/; see In'nami et al., 2022, for instructions on creating simulated data).

Author statement

There are no conflicts of interest to declare.

Acknowledgements

We are grateful to Kazuhiko Katagiri, Ikuko Koshimizu, Makoto Fukazawa, Yumi Koyamada, and the Eiken Foundation of Japan for their constant assistance with this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.system.2023.103208.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Boersma, P., & Weenink, D. (2023). Praat: Doing phonetics by computer [online software] https://www.fon.hum.uva.nl/praat/.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition, 40*(2), 443–457. https://doi.org/10.1017/S027226311700016X

De Jong, N. H. (2023). Assessing second language speaking proficiency. *Annual Review of Linguistics, 9*, 541–560. https://doi.org/10.1146/annurev-linguistics-030521-052114

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

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012b). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

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), 387–404. https://doi.org/10.1177/1362168815606161

Ekiert, M., Révész, A., Torgersen, E., & Moss, E. (2022). The role of pausing in L2 oral task performance. *Journal on Task-Based Language Teaching and Learning, 2*(1), 33–59. https://doi.org/10.1075/task.21013.eki

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE.

Foster, P. (2020). Oral fluency in a second language: A research agenda for the next ten years. *Language Teaching, 53*(4), 446–461. https://doi.org/10.1017/S026144482000018X

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

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics, 36*, 98–116. https://doi.org/10.1017/S0267190515000082

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). John Benjamins.

In'nami, Y., Mizumoto, A., Plonsky, L., & Koizumi, R. (2022). Promoting computationally reproducible research in applied linguistics: Recommended practices and considerations. *Research Methods in Applied Linguistics, 1*(3), 1–9. https://doi.org/10.1016/j.rmal.2022.100030, 100030.

Inoue, C. (2011). *Task parallelness: Investigating the difficulty of two spoken narrative tasks [doctoral dissertation*. Lancaster University. https://eprints.lancs.ac.uk/id/eprint/133463/.

Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 570–584). Routledge.

Kakitani, J. (2023). Equivalency of picture-based speaking tasks: An investigation of complexity, accuracy, lexis, and fluency. *The Language Teacher, 47*(2), 3–10. https://doi.org/10.37546/JALTTLT47.2-1

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox.

Koizumi, R., & In'nami, Y. (2022). Assessing functional adequacy using picture description tasks in classroom-based L2 speaking assessment. *JLTA Journal, 25*, 60–79. https://doi.org/10.20622/jltajournal.25.0_60.

Koizumi, R., In'nami, Y., & Jeon, E. H. (2022). L2 speaking and its internal correlates: A meta-analysis. In E. H. Jeon, & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 307–338). John Benjamins. https://doi.org/10.1075/bpa.13.

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

Kuiken, F., & Vedder, I. (2018). Assessing functional adequacy of L2 performance in a task-based approach. In N. Taguchi, & Y. J. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 265–285). John Benjamins.

Kuiken, F., & Vedder, I. (2022a). The assessment of functional adequacy in language performance. *Journal on Task-Based Language Teaching and Learning, 2*(1), 1–7. https://doi.org/10.1075/task.21009.kui

Kuiken, F., & Vedder, I. (2022b). Measurement of functional adequacy in different learning contexts: Rationale, key issues, and future perspectives. *Journal on Task-Based Language Teaching and Learning, 2*(1), 8–32. https://doi.org/10.1075/task.00013.kui

Kuiken, F., & Vedder, I. (2022c). Speaking: Complexity, accuracy, fluency, and functional adequacy (CAFFA). In L. Gurzynski-Weiss, & Y. J. Kim (Eds.), *Instructed second language acquisition research methods* (pp. 329–352). John Benjamins. https://doi.org/10.1075/rmal.3.

Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research (eurosla monographs 1* (pp. 81–99). European Second Language Acquisition http://eurosla.org/monographs/EM01/EM01home.html.

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

Pallotti, G. (2009). Caf: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Pallotti, G. (2022). Holistic and analytic assessment of functional adequacy. *Journal on Task-Based Language Teaching and Learning, 2*(1), 85–114. https://doi.org/10.1075/task.21014.pal

Révész, A., Ekiert, M., & Torgersen, E. N. (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

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching, 43*(1), 1–32. https://doi.org/10.1515/iral.2005.43.1.1

Rossiter, M. J., Derwing, T. M., & Jones, V. M. L. O. (2008). Is a picture worth a thousand words? *TESOL Quarterly, 42*(2), 325–329. https://doi.org/10.1002/j.1545-7249.2008.tb00127.x

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*(2), 217–240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38*(4), 439–462. https://doi.org/10.1093/applin/amv047

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

Strobl, C., & Baten, K. (2022). Assessing writing development during study abroad. *Journal on Task-Based Language Teaching and Learning, 2*(1), 60–84. https://doi.org/10.1075/task.21010.str

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

Suzuki, Y., & Koizumi, R. (2021). Using equivalent test forms in SLA pretest-posttest design research. In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 457–467). Routledge.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47*, 1–15. https://doi.org/10.1016/j.asw.2020.100505. Article 100505.