What shapes communicative adequacy in second language speaking performance? The contributions of complexity, accuracy, fluency, and pronunciation

Zhuo Chen
Department of General Courses
Guangzhou Panyu Polytechnic, China
zhuochen@m.scnu.edu.cn

Abstract

The necessity of considering communicative adequacy (CA) in assessing second language (L2) performance has been increasingly recognized, while its nature has yet to be fully explored. The present study examines the relationship between CA and the dimensions of complexity, accuracy, fluency, and pronunciation (CAFP) in L2 speaking assessment. Specifically, the speaking performance of 158 Chinese learners of English was subjectively rated in terms of CA and was also subjectively rated and objectively measured in CAFP. The relationship between the subjective ratings of CA and CAFP and the relationship between the subjective ratings of CA and CAFP and the objective measures of CAFP were analyzed. Results show that the subjective ratings of all CAFP dimensions were significantly correlated with and predicted CA, with pronunciation and fluency ratings making relatively greater contributions to CA than complexity and accuracy ratings, while only the objective measures of verbal complexity, speed fluency, and pronunciation significantly correlated with CA, together accounting for 45% of CA's variance. Furthermore, the subjective ratings of CAFP showed limited correlations with their objective measures. Discussions were made concerning the validity of the construct of CA, the relative contributions of CAFP to CA, and the important role of pronunciation in L2 speaking assessment.

Keywords: communicative adequacy; complexity; accuracy; fluency; pronunciation.

Resumen

La necesidad de considerar la adecuación comunicativa (CA) en la evaluación del desempeño en el segundo idioma (L2) ha sido reconocida cada vez más, aunque su naturaleza aún no ha sido completamente explorada. El presente estudio examina la relación entre la CA y las dimensiones de complejidad, precisión, fluidez y pronunciación (CAFP) en la evaluación del habla en la L2. Específicamente, el desempeño oral de 158 estudiantes chinos de inglés fue evaluado subjetivamente en términos de CA, y también fue evaluado subjetiva y objetivamente en CAFP. Se analizó la relación entre las evaluaciones subjetivas de CA y CAFP, así como la relación entre las evaluaciones subjetivas de CA y CAFP y las medidas objetivas de CAFP. Los resultados muestran que las evaluaciones subjetivas de todas las dimensiones de CAFP están significativamente correlacionadas y predicen la CA, siendo las evaluaciones de pronunciación y fluidez las que contribuyen relativamente más a la CA que la complejidad y precisión, mientras que solo las medidas objetivas de complejidad verbal, fluidez y pronunciación se correlacionaron significativamente con la CA, representando conjuntamente el 45% de la varianza de la CA. Además, las evaluaciones subjetivas de CAFP mostraron correlaciones limitadas con sus medidas objetivas. Se realizaron discusiones sobre la validez del constructo de CA, las contribuciones relativas de CAFP a la CA y el papel importante de la pronunciación en la evaluación del habla en la L2.

Palabras clave: adecuación comunicativa; complejidad; precisión; fluidez; pronunciación.

1. Introduction

Communicative adequacy (CA) represents how successful a second language (L2) performance achieves its communicative goals (Pallotti, 2009). It has been increasingly used as a way to assess L2 speaking performance alongside other traditional measures such as accuracy and fluency (Koizumi & In'nami, 2024; Kuiken & Vedder, 2022b, 2022c; Révész et al., 2016). This is not only because adequacy is a major goal for L2 learners (Kuiken & Vedder, 2022a) and a basic requirement for effective oral communication (De Jong, 2023) but also for its complementary role to the complexity, accuracy, and fluency (CAF) triad (Pallotti, 2021). However, little is known about the nature of CA, i.e., what linguistic features and to what extent different linguistic features contribute to the CA of L2 performance, and how reliable it could be judged. One way to address these issues is to relate CA with other well-established constructs, as in Révész et al. (2016). Indeed, researchers have long been relating different assessments, subjective or objective, global or specific, of the same language performance to address various issues, for example, to explore the nature of certain assessment methods (Hulstijn et al., 2012; Plakans et al., 2019), the structure of L2 proficiency (Bosker et al., 2013; Suzuki & Kormos, 2020), and the raters' rating process (Kormos & Dénes, 2004; Kuiken & Vedder, 2014a).

The first appropriate benchmark for CA is CAF (Housen et al., 2012; Pallotti, 2021), since it has been most widely used and researched in language testing as well as theoretically and empirically justified as a valid and reliable construct to gauge L2 performance and proficiency (Larsen-Freeman, 2006; Skehan, 2003). Nonetheless, it is arguable that CAF, which focuses more on the aspects of grammar, would not be sufficient to test CA without the dimension of pronunciation which captures the phonetic features. To begin with, CAF has been challenged for its comprehensiveness (Skehan, 2009). Moreover, pronunciation is a salient and basic dimension of L2 speaking performance (De Jong & Van Ginkel, 1992; De Jong et al., 2012a) and an integral part of major speaking assessments, and thus warrants to be weighed alongside CAF. Besides, pronunciation has been found to influence the comprehensibility (i.e., ease of understanding) of L2 oral production (Suzuki & Kormos, 2020) and impact the success of communication or CA (Derwing & Munro, 2009).

As such, in order to explore the validity of the construct of CA and the relative importance of CAFP in assessing CA, the present study investigates the relationship between three types of assessment of the same speaking performances, namely, raters' subjective ratings and several objective measures of the CAFP of these performances, as well as raters' subjective ratings of their CA. It is hoped that the results can offer insights into the theoretical construction and the pedagogical practice of L2 speaking assessment.

2. Literature review

2.1. Communicative adequacy

Pallotti (2009) was one of the first to explicitly address the importance of adequacy in relation to CAF. Based on a scrutiny of the major problems in defining and operationalizing the constructs of CAF, he proposed that *adequacy*, which is defined as "the degree to which a learners' performance is more or less successful in achieving the task's goals efficiently" (p. 596), should be used as an independent construct, complementing and interpreting measures of CAF. Since Pallotti's seminal work, the argument that interpretations of the CAF of L2 performance may not be satisfactory without considering adequacy has been increasingly recognized (De Jong et al., 2012b; Kuiken & Vedder, 2022a). Several empirical studies have been carried out to explore the nature of the construct of CA and its relationship with CAF.

Hulstijn et al. (2012) analyzed the relationship between L2 Dutch learners' speaking proficiency rated in terms of communicative adequacy, i.e., "the adequacy with which participants were able to perform communicative speaking tasks" (p. 205) and several other aspects of their linguistic competences. They extrapolated the CA scores into the B1 or B2 speaking proficiency levels of the Common European Framework of Reference for Languages and found that different linguistic competences had varied differentiating power to the two proficiency levels, thus implicitly demonstrating the usefulness of CA in assessing L2 speaking performance. Drawing from the same data pool, De Jong et al. (2012a) directly related L2 learners' functional adequacy (a measure of communicative success) to their linguistic skills. Analyses revealed that, except for the speed of articulation measures (i.e., response latency and response duration), all the other skills were significantly related to the learners' functional adequacy scores, together accounting for 76% of its variance. This result illustrates the componential nature of functional adequacy. De Jong et al. (2012b) further demonstrated that different degrees of task complexity (complex vs. simple tasks) influence functional adequacy measures (rated on six levels), fluency measures (including breakdown fluency, speed fluency, and repair fluency), and lexical diversity measures (Guiraud's index) in different ways. This result effectively supports their argument that measuring CAF "does not amount to the same thing as measuring overall speaking performance" (p. 135). They suggested that CAF combined with CA is likely to predict the overall success of a speaking performance.

In the area of L2 writing assessment, using the data from the Communicative Adequacy and Linguistic Complexity (CALC) study, Kuiken et al. (2010), and Kuiken and Vedder (2014b) looked specifically into the construct of CA and addressed its relationship with linguistic complexity. Kuiken et al. (2010) found significant high correlations between the subjective ratings of CA and the subjective ratings of linguistic complexity. As for the objective measures of complexity, significant correlations were found between CA and lexical diversity and accuracy, but not with syntactic complexity. Kuiken and Vedder (2014b) further revealed that significant correlations existed between overall ratings of CA and linguistic complexity in L2 as well as L1 writing assessment. In addition, raters reported attaching more importance to communicative adequacy than to linguistic complexity. The two studies have expanded our understanding of the nature of CA in L2 writing assessment. More importantly, they justified the research effort of making distinctions between CA and CAF and demonstrated the usefulness of the construct of CA in L2 assessment.

It is worth noting that there is a discrepancy in terminology in the studies presented in this subsection so far. Specifically, half of these studies used the term *communicative adequacy*, while the other half used *functional adequacy.* Kuiken et al. (2010), and Kuiken and Vedder (2014b) used these two different terms for the same construct, which was explicitly defined as "how well participants manage to fulfill the communicative requirements set by the speaking task" (De Jong et al., 2012b, p. 123) based on Pallotti (2009). To maintain consistency, the present study employs the term *communicative adequacy*, following Pallotti (2009, p. 599), who referred to adequacy as "the appropriateness to communicative goals and situations", considering that *communicative adequacy* displays a focus on whether an L2 speaking performance succeeds in achieving the intended communicative goal. Additionally, this is to distinguish it from the task-related construct of *functional adequacy* that Kuiken and Vedder (2017) developed for L2 writing production.

Taken together, previous studies on CA featured a general agreement on Pallotti's (2009) basic argument of CA and its relative role to CAF, but a coherent and clear definition of CA is still lacking, showing a limited understanding of its nature. Besides, the objective measures of the three dimensions of CAF can only shed light on our understanding of the nature of the construct of CA, and these measures alone are not enough for a comprehensive understanding. Moreover, one or two linguistic dimensions of complexity or fluency would not be sufficient to present a holistic construct as CA and there is still no empirical evidence to show that measuring the three dimensions of CAF jointly equals measuring CA. Therefore, exploring the relationship of CA with CAF is "the most tempting endeavor" (Kuiken et al., 2010, p. 95).

Révész et al. (2016) conducted a study to this aim. They investigated the extent to which CAF predicted CA in L2 oral tasks. Results showed that filled pause frequency was the strongest predictor of CA, explaining 15% of CA's variability, while all other indices of CAF dimensions had significant but small contributions (from 1% to 7%). A model with all the objective measures as predictors accounted for 41% of CA's variance, a power much weaker than the 76% in De Jong et al. (2012a). Their main explanation was that they did not take pronunciation quality as a variable, while in De Jong et al. (2012a), a pronunciation measure was included and showed the strongest contribution (34%) among all the nine linguistic features. Specifically, Révész et al. (2016, p. 846) stated that "A possible explanation why our model was able to explain less variance may lie in the fact that we did not consider pronunciation quality, while this factor had a strong impact on adequacy in De Jong et al. (2012a)". According to their statement, "pronunciation quality" is similar to the "pronunciation quality" analyzed in De Jong et al. (2012a, p. 11), which includes the correctness of speech sounds, word stress, and intonation. Moreover, according to them, it would have been possible to investigate pronunciation in their dataset as De Jong et al. (2012a) did. Whether what Révész et al. (2016) explained was the case should be examined empirically, as it has significant implications for L2 teaching and testing practices regarding the role of pronunciation.

2.2. Complexity, accuracy, fluency, and pronunciation

Unlike CA, the CAF triad has been explored extensively in SLA research (Pallotti, 2021). Most researchers agree that the multifaceted and multidimensional nature of L2 language performance and proficiency is most frequently and adequately captured by the constructs of CAF (e.g., Housen & Kuiken, 2009; Skehan, 2003). In this study, a fourth dimension - pronunciation is added to CAF in response to the special nature of oral performance and the argument that pronunciation is an integral part of L2 speaking proficiency (Isaacs & Thomson, 2013; Suzuki & Kormos, 2020), resulting in the quartet of CAFP. This addition was also made for the following reasons.

Firstly, although the construct of pronunciation is operationalized differently in different studies, they all focus on the phonological and the acoustic aspects of oral language production, and it has been shown to play an important role in judging overall spoken language proficiency. For example, Higgs and Clifford (1982) found that pronunciation was important across different proficiency levels of the FSI (Foreign Service Institute) speaking test scale. In De Jong and Van Ginkel (1992), pronunciation contributed most to a global proficiency score at lower proficiency levels, and at higher levels it made equal contributions as fluency and accuracy. Iwashita et al. (2008) also revealed that pronunciation in terms of L1 target-like syllables was showing greater relative impact on overall proficiency scores.

Secondly, studies have shown that pronunciation figured prominently in the comprehensibility of L2 speeches, which is "central to interlocutors' communicative success" (Isaacs & Trofimovich, 2012, p. 475). Therefore, it is particularly pertinent to our analysis of the contributing factors of CA. Isaacs and Trofimovich (2012) examined 19 segmental, suprasegmental, fluency, lexical, grammatical, and discourse-level variables to explore which linguistic features influence raters' judgments of L2 comprehensibility at different proficiency levels. The pronunciation measure was the only one showing significant discriminating power between all levels of learners. Saito and Shintani (2016) also found that raters' judgment of L2 speech comprehensibility was primarily influenced by pronunciation and fluency measures and secondarily by lexical and grammatical ones. Similarly, in Suzuki & Kormos (2020), raters' judgment of comprehensibility was best predicted by the objective measures of fluency, grammatical accuracy, and pronunciation. Moreover, in the post-rating interview, pronunciation was the only feature mentioned by all the raters to have influenced their judgment of comprehensibility.

In addition, pronunciation has been found to be independent of CAF. De Jong and Van Ginkel (1992) revealed that accuracy and comprehensibility were two perspectives of a single dimension of message conveyance, while pronunciation and fluency constitute two separate dimensions of speech production. Therefore, they suggested that accuracy, pronunciation, and fluency be treated and measured separately. In Pinget et al. (2014), raters rated the foreign accentedness of the speeches on a nine-point scale ranging from 1 'no accent' to 9 'very strong accent' based on their judgments on the pronunciation of sounds, word stress and intonation patterns.

Their ratings show how much the speakers' pronunciation deviated from the norms of Standard Dutch (p. 359). They also found that raters' ratings of fluency and perceived foreign accent were predicted by different objective acoustic measures and they were only weakly correlated. This finding supported their argument that fluency and accent were separate constructs and should be assessed independently. Note that the raters' rating of accent was based on their judgment of the pronunciation of sounds, word stress, and intonation, and the objective measures of accent consisted of phonemic error rate and pitch range. Therefore, their result for the variable of accent was also indicative of the interdependent relationship between fluency and pronunciation.

3. Research questions

As such, the present study was designed to address the following research questions (RQs).

RQ1: How do raters' subjective ratings of the CAFP of the test takers' speaking performance relate to their subjective ratings of the CA of the same performance?

RQ2: How do the objective measures of the CAFP of the test takers' speaking performance relate to raters' subjective ratings of the CA of the same performance?

RQ3: How do the objective measures of the CAFP of the test takers' speaking performance relate to raters' subjective ratings of the CAFP of the same performance?

4. Methodology

4.1. Participants

One hundred and fifty-eight year-one postgraduate students in an engineering university in China took part in the study on a voluntary basis. They all passed the national entrance examination for postgraduate students in China, which included an English test. According to the examination syllabus, their English proficiency was intermediate to advanced. Their ages ranged from 22 to 28 (*M* = 22.59, *SD* = 1.80), with 53 females and 105 males.

4.2. Speaking task

The subjects were instructed to make a 2-3 minutes' speech on the topic of "A positive change in life". They had two minutes to prepare for the speech. Their speeches were rated on the spot for communicative adequacy and they were also recorded for further rating, coding, and measurement.

4.3. Transcription

The 158 recorded speeches were 0.73-3.57 minutes long (*M* = 1.89, *SD* = 0.56) and were transcribed by an experienced researcher in applied linguistics. Twenty randomly selected transcripts were checked for accuracy by another experienced researcher in applied linguistics to ensure reliability. The check reveals that the first researcher's transcription was accurate and reliable.

4.4. Assessment

4.4.1. Raters' subjective ratings of CA and CAFP

Eight experienced raters, who were also EFL teachers in the subject university, were employed to ensure the ecological validity of the study (Kormos & Dénes, 2004). However, the teacher raters might have been biased in what to listen for, as they were experienced EFL teachers and were familiar with the common mistakes in students' oral performance. To reduce the possibility of any bias, they were clearly informed of the criteria for rating and were given trials to ensure that they fully understood the criteria before they rated the target speeches.

The raters rated the speeches in two ways, i.e., holistically in terms of CA and specifically in CAFP. The holistic rating of CA was done on the spot during the test by the eight raters. They were provided with the definition by Pallotti (2009) (presented in the Literature Review). Following the practice in De Jong et al. (2012a), Hulstijn et al. (2012), Kuiken and Vender (2014b), and others, they were asked to rate the CA of the participants' performance on the same 6-level scales (i.e., *unsuccessful*, *weak*, *mediocre*, *sufficient*, *quite successful*, and *very successful*), but with no detailed descriptors, to ensure that the raters made an intuitive judgment of this insufficiently researched construct (Suzuki & Kormos, 2020). Another consideration is that, as De Jong (2018) pointed out, for the studies relating measures of CAF to overall proficiency, one serious problem is that "what is apparent in the descriptors is likely to emerge as a significant predictor" (p. 243). For example, if the rubrics of oral proficiency contain descriptions of fluency, then it would be natural that a moderate or strong correlation would be found between fluency measures and proficiency ratings (e.g., in Iwashita et al., 2008). Therefore, we purposefully chose not to provide detailed rubrics for the rating of CA. This was also employed to check whether as a construct proposed by SLA researchers, CA is sensible and valid for other practitioners. Nonetheless, the inter-rater reliability coefficient in terms of Cronbach's Alpha for CA was 0.933, showing that the raters were consistent in their judgment. Their average was calculated as the final score for each participant.

The rating of CAFP was done by the same raters based on the recordings of the participants' speeches three weeks after the test. The raters were not informed before the second rating task to ensure that they were not biased or influenced by their first holistic rating practice. Unlike the rating of CA, the rating of CAFP was guided by rubrics adopted from the IELTS speaking band descriptors to ensure that the raters treated CAFP separately, not treating any of them in a broad sense (see Bosker et al., 2013, for discussions of fluency broadly defined being interpreted as overall speaking proficiency). Besides, we used different rating scales for the two ratings, i.e., a 6-level scale for CA and a 9-point scale for CAFP, to avoid artifactual test homogeneity (De Jong & Van Ginkel, 1992) and to further reduce possible interference from the first rating practice. The Cronbach's Alpha for the raters' rating of CAFP was 0.969, 0.946, 0.973, and 0.972, respectively, also showing a high degree of consistency. Similarly, their scores of CAFP were averaged to obtain a final score for each subject.

It should also be noted that unlike Bosker et al. (2013), who used different groups of raters for overall and specific ratings, we deem it important to use the same raters for holistic and specific ratings, as the possible consistency or inconsistency between the ratings of CA and CAFP by the same raters may better reveal the nature of CA, while using different groups of raters is itself a potential source of discrepancy.

After the second rating task, the raters were also asked to rank the relative importance of CAFP in CA.

4.4.2. Objective measures of CAFP

The operationalization of CAFP followed three major principles. First, we employed the widely accepted definitions and measures in the literature to ensure comparability with similar studies. Second, we selected the measures that are generally found to be effective to ensure their validity. Most importantly, instead of trying to achieve full coverage, we used limited yet complementary and distinct measures for each construct to avoid metrics redundancy (Norris & Ortega, 2009) and to reduce the chance of confounding different measures and leading to intercollinearity, since it would make it impossible to estimate the variance uniquely predicted by each predictor variable (De Jong, 2018).

Complexity. Complexity has been variously defined, for example, as "the extent to which learners produce elaborated language" (Ellis & Barkhuizen, 2005, p. 139) or as "the size, elaborateness, richness and diversity of the learner's linguistic L2 system" (Housen & Kuiken, 2009, p. 465). According to Pallotti (2009), complexity poses most problems among the CAF triad for its polysemous nature. For instance, Ellis (2003) listed as many as eight measures of complexity.

This study employed a generally used measure – the amount of subordination, specifically, the ratio of clauses to the Analysis of Speech Units (AS-units) in the subjects' speeches. It was complemented by a measure of specific linguistic features, i.e., the number of different grammatical verb forms in their speeches, including tense, modality, and voice. AS-unit was used instead of t-unit to better suit the features of oral texts (Foster et al., 2000).

Accuracy. Accuracy is generally constructed as "the ability to produce error-free speech" (Housen & Kuiken, 2009, p. 461). The two most often identified measures in the literature are general accuracy and specific measures of accuracy. The former refers to identifying all types of errors and calculating the percentage of error-free clauses or errors per 100 words (e.g., Skehan & Foster, 1999) or the incidence of errors per t-unit (e.g., Bygate et al., 2013), while the latter refers to identifying particular types of error, for example, correct use of vocabulary (Skehan & Foster, 1997) and correct use of plurals (Ortega, 1999).

Following Ellis and Barkhuizen's (2005) proposal, the general measure of accuracy used in this study was "errors per 100 words" and the specific measure was "correct verb forms", which has been long and widely used in the literature (Wigglesworth, 1997; Yuan & Ellis, 2003) (See Table 1).

Fluency. Fluency has also been defined in a variety of ways, for example, as "the capacity to produce speech at normal rate and without interruption" (Skehan, 2009, p. 510), or "the production of language in real time without undue pausing or hesitation" (Ellis & Barkhuizen, 2005, p. 139). Its measurements are mostly of two kinds: in terms of temporal variable, being related to the speed of speaking/writing, and by means of hesitation phenomena, being related to pauses, repetitions, and other dysfluency features (Lennon, 1990).

The present study also operationalized fluency in these two ways. One is the speed of speaking, measured by the number of syllables per minute of speech (Yuan & Ellis, 2003). The other is dysfluency or repair fluency, indexed by the number of reformulations, repetitions, false starts, and replacements per AS-unit, the definitions of which followed Skehan and Foster (1999).

Pronunciation. Two major ways of measuring pronunciation can be identified from the literature. First, the subjects are instructed to read given linguistic materials which include some target words or sounds. The reading is then analyzed with only the chosen words or sounds assessed (e.g., in De Jong et al., 2012a). In the second approach, the subjects' elicited speech is analyzed for certain segmental or suprasegmental features, for instance, how target-like the pronunciation of meaningful words and syllables was

(Iwashita et al., 2008, p. 33), or by calculating the segmental, word stress, or supersegmental error rate (e.g., in Suzuki & Kormos, 2020). Our pronunciation measure mainly referred to Saito and Shintani (2016) and considered the number of erroneously pronounced words, that is, words that have segmental errors (substitution, omission, or insertion of individual consonant and vowel sounds) and word stress errors (misplaced or missing primary stress). The pronunciation was compared to standard British English and American English pronunciation, as that is what the subjects had been learning at school.

All measures and calculations of CAFP are summarized in Table 1.

Table 1: Summary of objective CAFP measures and their calculations

Dimensions	Aspects	Calculations
Complexity	Syntactic	Ratio of clauses to AS-units
	Verbal	Number of different grammatical verb forms in the speech
Accuracy	Specific	Percentage of all verbs used correctly to the total number of verbs
	General	Errors per 100 words
Fluency	Speed	Number of syllables per minute of speech
	Repair	Number of reformulations, repetitions, false starts, and replacements per AS-unit
Pronunciation	–	Number of erroneously pronounced words per 100 words

The first researcher analyzed the transcripts and made the computations. Twenty randomly selected transcripts were coded by the second researcher. The inter-rater correlation coefficient was 0.732, indicating a moderate and acceptable level of reliability (Koo & Li, 2016).

4.5. Data analysis

Three sessions of analyses were carried out using r software (R Core Team, 2023) to address the research questions, namely, descriptive, correlation, and regression analyses. The statistical analysis was complemented by the raters' view of the relative importance of CAFP in CA.

5. Results

5.1. Descriptive statistics

The descriptive statistics of subjectively rated and objectively measured CAFP and subjectively rated CA are summarized in Table 2. Skewness and kurtosis values verify that the scores were normally distributed. Mean scores of subjectively rated CAFP suggest that the subjects had relatively high scores in accuracy, followed in turn by fluency, pronunciation, and complexity.

Table 2: Descriptive statistics of CAFP and CA measures (*N* = 158)

Types	Measures	Mean	SD
Subjective	Complexity	5.521	0.727
	Accuracy	5.730	0.669
	Fluency	5.619	0.879
	Pronunciation	5.623	0.697
	CAFP average	5.624	0.626
	CA	3.960	0.501
Objective	Syntactic complexity	1.646	0.254
	Verbal complexity	4.171	1.365
	Specific accuracy	0.807	0.093
	General accuracy	5.427	2.485
	Speed fluency	118.704	28.230
	Repair fluency	5.672	4.358
	Pronunciation	1.075	1.171

5.2. Results of Research Question 1

The first research question explores how raters' subjective ratings of the specific dimensions of CAFP relate to their subjective ratings of CA. It was answered through a series of correlation and regression analyses. The results of Pearson correlation analysis between the raters' ratings of CAFP and between the ratings of CAFP and CA are presented in Table 3.

Table 3: Correlation between subjective ratings of CAFP and between subjective ratings of CAFP and CA (*N* = 158)

	Complexity	Accuracy	Fluency	Pronunciation	CAFP average
Complexity					
Accuracy	0.533**				
Fluency	0.649**	0.580**			
Pronunciation	0.688**	0.595**	0.614**		
CAFP average	0.852**	0.791**	0.865**	0.853**	
CA	0.764**	0.745**	0.793**	0.857**	0.938**

Note. ***p* < .01.

The correlation coefficients between the ratings of CAFP range from 0.533 to 0.688 (p < .01), representing an interrelated yet interdependent relationship among these four dimensions. As for the subjective ratings of CA and CAFP, although the two types of assessments were conducted separately with an interval of three weeks, they were strongly correlated (r = 0.745 - 0.857, p < .01), especially for the average score of CAFP and the rating of CA (r = 0.938, p < .01), indicating that the raters were consistent in rating and their judgment of L2 oral performance. Among CAFP, pronunciation and fluency had relatively higher correlations with CA than complexity and accuracy.

To investigate the extent to which CAFP accounts for the variance of the ratings of CA, a set of stepwise regression analyses were conducted. First, each variable of CAFP was used to predict CA separately. The values of adjusted R^2 shown in Table 4 indicate that pronunciation is the strongest predictor of CA, explaining 73% of the observed variance in CA, with fluency, complexity, and accuracy accounting for 63%, 58%, and 55%, respectively (p < .001).

Table 4: Results of regression analyses for subjective ratings of CAFP as predictors of CA

Predictors							
Step 1	R^2	**Adjusted** R^2	**Estimates**	*SE*	**Statistic**	p	**VIF**
Pronunciation	0.734	0.732	0.62	0.3	20.75	<0.001	2.29
Fluency	0.629	0.626	0.45	0.03	16.26	<0.001	2.07
Complexity	0.583	0.581	0.53	0.04	14.78	<0.001	2.27
Accuracy	0.555	0.552	0.56	0.04	13.94	<0.001	1.76
Step 2							
P + F + C+A	0.894	0.891					
Pronunciation (P)			0.31	0.03	10.88	<0.001	
Fluency (F)			0.17	0.02	7.74	<0.001	
Complexity (C)			0.10	0.03	3.76	<0.001	
Accuracy (A)			0.18	0.03	6.82	<0.001	

Secondly, we fitted multilevel models with more than one variable as predictors. According to Plonsky and Ghanbar (2018), the assumptions of multiple regression, e.g., normality, outliers, and multi-collinearity among predictor variables, were checked. Shapiro-Wilk normality tests were carried out on all variables, and the *p* values, ranging from 0.622 to 0.838, confirmed their normal distribution. Variance Inflation Rate (VIF) values were calculated to check multi-collinearity, which showed no such problem (VIF = 1.76 - 2.29). The best fitted model, with the lowest AIC (Akaike Information Criterion) of -113.81, has pronunciation, fluency, accuracy, and complexity as predictors, accounting for 89% of the variance of CA (see Table 4).

Additionally, the results of the raters' view on the relative importance of CAFP in rating CA show that, on average, the eight raters took accuracy and pronunciation as equally important, followed by fluency, and they all ranked complexity as the least important.

5.3. Results of Research Question 2

The second research question addresses how objective measures of CAFP relate to raters' subjective ratings of CA. Similar correlation and regression analyses were carried out.

Table 5 shows few significant correlations among the objective measures of CAFP. Within each of the four dimensions, only general accuracy significantly correlated with specific accuracy, while the two measures of complexity had no significant correlation with each other, nor did the two measures of fluency. As for the relationship between the objective measures of CAFP and the subjective ratings of CA, speed fluency and verbal complexity moderately correlated with the subjective ratings of CA, with objective measures and subjective ratings of pronunciation showing a significant yet limited correlation.

Table 5: Correlation between objective measures of CAFP and between objective measures of CAFP and subjective ratings of CA (*N* = 158)

	Syntactic complexity	Verbal complexity	Specific accuracy	General accuracy	Speed fluency	Repair fluency	Pronunciation
Syntactic complexity							
Verbal complexity	0.015						
Specific accuracy	0.174*	-0.149					
General accuracy	0.145	0.060	-0.724**				
Speed fluency	-0.143	0.281**	0.080	-0.056			
Repair fluency	-0.472**	0.096	-0.160*	-0.238**	-0.262		
Pronunciation	0.283	-0.131	0.011	0.147	0.019	-0.343**	
CA	-0.010	0.478**	-0.216	0.198	0.513**	-0.148	-0.303*

Note. **p* < .05, ***p* < .01.

Each objective measure was then used to predict CA. The adjusted R^2 presented in Table 6 shows that speed fluency, verbal complexity, and pronunciation accounted for 26%, 23%, and 9% of CA's variance, respectively.

Besides, models with more than one variable as predictors were also fitted to find the best fitting model. Preliminaries such as normality, outliers, and multi-collinearity among predictor variables were also checked and no such problems were revealed. As in Suzuki and Kormos (2020), and Saito and Shintani (2016), the linguistic measures that did not correlate significantly with the outcome variables were excluded from constructing the model. Subsequently, speed fluency, verbal complexity, and pronunciation together significantly predicted the raters' subjective judgment of CA, explaining 45% of its variance.

Table 6: Results of regression analyses for objective measures of CAFP as predictors of subjective ratings of CA

Predictors					
Step 1	***R*²**	Adjusted ***R*²**	Estimates	***SE***	***p***
Speed fluency	0.263	0.258	0.01	0.00	<0.001
Verbal complexity	0.229	0.224	0.18	0.03	<0.001
Pronunciation	0.092	0.086	-0.13	0.03	<0.001
Specific accuracy	0.047	0.041	-1.17	0.42	0.053
General accuracy	0.039	0.033	0.04	0.02	0.056
Repair fluency	0.022	0.016	-0.02	0.01	0.063
Syntactic complexity	0.000	-0.006	-0.02	0.16	0.898
Step 2					
SF + VC + P	0.455	0.445			
Speed fluency (SF)			0.01	0.00	<0.001
Verbal complexity (VC)			0.12	0.02	<0.001
Pronunciation (P)			-0.12	0.03	<0.001

5.4. Results of Research Question 3

Table 7: Correlation between objective measures and subjective ratings of CAFP (*N* = 158)

Objective measures	**Subjective ratings**			
	Complexity	**Accuracy**	**Fluency**	**Pronunciation**
Syntactic complexity	0.011	-0.060	0.237**	-0.128
Verbal complexity	0.568**	0.182*	0.350	0.449
Specific accuracy	-0.158*	-0.126	-0.114	-0.236**
General accuracy	0.148	0.138	0.286**	0.183*
Speed fluency	0.399**	0.540**	0.504**	0.353**
Repair fluency	-0.103	-0.162*	-0.494**	0.037
Pronunciation	-0.167*	-0.013	-0.032	-0.493**

Note. **p* < .05, ***p* < .01.

The third research question asks how the objective measures and subjective ratings of CAFP relate to each other. The Pearson correlation results summarized in Table 7 show that there was a moderate correlation between verbal complexity, speed fluency and repair fluency, and pronunciation with their corresponding subjective ratings.

Three regression models were accordingly constructed. Results in Table 8 show that objectively measured verbal complexity, fluency, and pronunciation accounted for 32%, 39%, and 24% of the variance of their corresponding subjective ratings.

Table 8: Results of regression analyses for objective measures of CAFP as predictors of their corresponding subjective ratings

Predictors	**Outcome variable**					
Objective	Subjective	R^2	Adjusted R^2	Estimates	*SE*	*p*
Verbal complexity	Complexity	0.323	0.318	0.30	0.04	<0.001
Pronunciation	Pronunciation	0.243	0.238	-0.29	0.04	<0.001
SF + RF	Fluency	0.394	0.387			
Speed fluency (SF)				0.01	0.00	<0.001
Repair fluency (RF)				-0.08	0.01	<0.001

6. Discussion

6.1. Discussion of Research Question 1

Results of the first research question show that subjectively rated CA strongly correlated with and was significantly predicted by all subjective ratings of CAFP. The best model fitted with all four variables accounted for 89% of the variance of CA. This adds to former research findings that objective measures of CAF were related to CA in L2 speaking (De Jong et al., 2012a; Révész et al., 2016) and that subjective ratings of linguistic complexity and accuracy were correlated with subjective ratings of CA in L2 writing (Kuiken & Vedder, 2014b).

Interestingly, unlike previous studies (Kuiken & Vender, 2014b; Révész et al., 2016), the present study employed very general descriptors of CA and still generated consistent results. Specifically, although, except for a general operational definition and basic descriptors of the 6-level scales of CA, no specifics were provided to the raters, the inter-rater reliability score was high for CA. It is even higher than that for CAFP, the rating of which was guided by detailed rubrics. This indicates that the raters seemed to agree on what CA means in oral performance and carried out the assessment with sufficient reliability. Therefore, in line with the former studies on CA, the present study enhances the evidence that CA is a valid construct for assessing L2 performance and could be reliably rated.

Another finding is that among the four dimensions of CAFP, the best predictors of CA were pronunciation and fluency, while the importance ranking data showed that the eight raters generally perceived accuracy and pronunciation more important for CA in L2 speaking than fluency and complexity. This interesting discrepancy between what raters did and what they thought may indicate that, metacognitively, raters believed that the L2 speech has to be accurate to be communicatively adequate. However, in actual communication, when the speech is above a threshold of accuracy and does not hinder understanding, its CA is not significantly affected (Pallotti, 2021). This finding bears out De Jong and Van Ginkel's (1992) view that oral production assessment is often based first on pronunciation and fluency, and then on the appropriateness of lexical or syntactical choices.

Comparing the present result with the results of other studies that examined the relative contribution of subjectively rated specific features to ratings of overall speaking proficiency, there are discrepancies as well as similarities. Our finding differs from McNamara (1990), who found a crucial role for grammatical and lexical accuracy but a limited role for fluency in assessing overall communicative effectiveness. Nonetheless, the present study partially replicates the findings of Higgs and Clifford (1982) that at lower proficiency levels, raters put most emphasis on vocabulary and pronunciation; as the level goes up, they gave more consideration to fluency and grammar. It is also in line with De Jong and Van Ginkel (1992) that at low proficiency levels, pronunciation contributed most to overall ability ratings, and then were accuracy and comprehensibility, with fluency contributing very little, while at higher proficiency levels, all subskills made equal contributions.

The similarities and discrepancies may, to some extent, be due to the fact that different studies chose different sets of dimensions and constructs, and even for the same dimensions, a wide range of measures and ratings were used.

The finding of the most significant role of pronunciation in predicting CA implies that when evaluating L2 speaking through specific performance features, pronunciation, as an important and unique feature of oral performance, should be adequately recognized and sufficiently analyzed (De Jong & Van Ginkel, 1992). This also suggests that the CAF triad may not present a comprehensive picture of L2 speaking performance (De Jong, 2023; Pallotti, 2009). However, even CAFP may not be comprehensive enough, as in this study, the four dimensions together accounted for 89% of the variance of CA. Since language performance cannot be meaningfully interpreted solely linguistically and without referring to its quality and effectiveness, we support the view of Pallotti (2009) and other relevant studies that communicative adequacy should be employed as "a separate performance dimension" and "as a way of interpreting CAF measures" (p. 590).

6.2. Discussion of Research Question 2

Among the objective measures of CAFP, only the two accuracy measures correlated with each other, while the two complexity measures did not, nor did the two fluency measures. Since we used "complementary and distinct measures" of CAFP to reduce the redundancy of measurements, it is reasonable that the measures of the same dimension had limited or no significant correlations.

Another main finding is that only three out of the seven objective measures of CAFP significantly predicted the subjective ratings of CA, i.e., verbal complexity, speed fluency, and pronunciation. This finding is not singular. In Révész et al. (2016), for nine out of ten CAF measures used to predict CA, their R^2 was quite low, ranging from 0.01 to 0.07. Only breakdown fluency made the highest contribution of 0.15. Similarly, among the nine linguistic skills analyzed by De Jong et al. (2012a), only knowledge of vocabulary and quality of intonation measures were significant predictors of CA, while all seven other variables made no significant contributions. In Douglas (1994), no significant relationship was observed between the subjective scores and objective measures. The underlying reason for this frequently reported limited relationship between the objective measures of CAF and the subjective ratings of oral proficiency may be that there exist a wide variety of measures for each CAF dimension. While each has its rationale, none or no combinations of them are sufficient enough to represent the whole of L2 oral proficiency, not to mention that the present study employed a holistic and more function-oriented construct as CA. However, this does not diminish the value of CAF in assessing the specific aspects of L2 performance and in describing its multidimensionality.

On the other hand, the significant role of verbal complexity, speed fluency, and pronunciation in predicting CA is also informative. Ellis and Barhuizen (2005) found that, unlike accuracy and fluency measures, complexity measures do not provide "a totally consistent picture" (p. 156), pointing to the fact that different complexity measures do not correlate closely with each other. Therefore, it is not unnatural that, in our study, verbal complexity had significant predictive power to CA, while syntactic complexity did not. Similarly, in L2 writing, Kuiken et al. (2010) found significant correlations between CA and lexical variation but not with syntactic complexity. The significant role of speed fluency and pronunciation replicates the results of several former studies. For instance, Iwashita et al. (2008) showed that token frequency, speech rate, and pronunciation significantly influenced oral language proficiency. Ginther et al. (2010) found that measures of speech rate had strong and moderate correlations with proficiency scores. In De Jong et al. (2012a), vocabulary knowledge and pronunciation were the best predictors of speaking proficiency.

The non-significant role of objective accuracy measures is also understandable. Note that the result of RQ1 shows that the contribution of subjective accuracy to CA is much lower than the other three dimensions. This may again reflect that, when rating CA, raters did not take accuracy as a primary influencing factor as long as the L2 production was accurate enough for understanding. Our result demonstrated Pallotti's (2009) illustration that an inaccurate statement like *No put green thing near bottle* is perfectly functional for achieving the intended communicative goal compared with *colorless green ideas sleep furiously*.

In all, it can be concluded from RQ2 that raters took a relatively holistic view when rating CA, which cannot be sufficiently captured by a combination of objective linguistic features. Additionally, different CAFP measures had different degrees of salience in predicting CA.

6.3. *Discussion of Research Question 3*

Under RQ3, the objective measures and subjective ratings of each CAFP dimension were related to each other. Similar results emerged and can help to explain the results of RQ1 and RQ2. Specifically, the objective measures of verbal complexity, fluency, and pronunciation significantly predicted their corresponding subjective ratings. The explanation of this limited significant relationship is similar to that of RQ2 and will not be repeated.

The significant relationship found between speed fluency and repair fluency with the subjective fluency ratings in this study has been reported in the literature. For example, Préfontaine et al. (2016) showed that the mean length of runs, articulation rate, and the frequency of pauses all played influential roles in raters' judgments of fluency. Besides, Kormos and Dénes (2004) also found that speech rate, mean length of runs, and pace significantly predicted subjective fluency scores. In Bosker et al. (2013), objective measures of pauses and speed were significant predictors of subjective fluency ratings. Taken together, it can be inferred that objective measures and subjective perceptions of the temporal features of oral production generally have a high degree of consistency.

7. Conclusion

7.1. General conclusion

This study explored the nature and the measurement of the construct of CA of L2 speaking performance by analyzing its relationship with the linguistic dimensions of CAFP. To achieve this, the oral English performance of 158 Chinese EFL learners was subjectively rated in terms of CAFP and CA and was objectively measured using seven indices of the CAFP dimensions. These three types of measurements were related to each other through a series of correlation and regression analyses. It was found that the subjective ratings of all CAFP dimensions are significantly correlated with and predicted CA, with pronunciation and fluency ratings making relatively greater contributions to CA than complexity and accuracy, while only the objective measures of verbal complexity, speed fluency, and pronunciation significantly correlated with CA. Furthermore, the subjective ratings of CAFP showed limited correlations with their objective measures. These findings point to the usefulness and validity of CA as a construct for assessing L2 speaking performance and they also demonstrate its complementary role to CAFP in L2 speaking assessment. Moreover, the significant role of pronunciation and fluency in predicting CA calls for more focused teaching and research effort. Furthermore, the limited correlation between the objective measures and subjective ratings highlights the importance of using both types of assessment to obtain a more comprehensive picture of L2 speaking performance.

7.2. Implications

The findings bear several implications. Theoretically, the construct of CA has been shown to have the potential to be taken as a promising measurement of L2 oral performance and it can sufficiently capture the facets of complexity, accuracy, fluency, and pronunciation of L2 speaking performance. Moreover, in assessing L2 speaking, the construct of CAFP is more valid and sufficient than CAF. Methodologically, this study adds to the empirical studies demonstrating the effectiveness of relating different types of assessment as a way to examine the assessment methods and explore the nature of L2 proficiency. Pedagogically, the prominent role of pronunciation in assessing oral performance calls for more teaching effort on pronunciation. Besides, the communicative or functional aspect of L2 speaking performance needs to be sufficiently considered in the teaching and testing practice. In addition, to get a comprehensive picture of L2 learners' speaking performance, it is recommended that both subjective ratings and objective measures be employed.

7.3. *Limitations*

One limitation of the study is that the result of the contributions of CAFP to CA was found with subjects who were at the intermediate to advanced levels of proficiency. According to the Relative Contributions Model (Higgs & Clifford, 1982), different proficiency levels would see different magnitudes of influence from the same linguistic factors. Therefore, it is worthwhile to examine the research questions with learners at lower proficiency levels. Moreover, the raters were only asked to rank the relative importance of CAFP for CA, but the underlying reasons or beliefs were left unexplored. Post-rating interviews would be more revealing (as in Trofimovich & Isaacs, 2012). Last, the analysis of CA in the literature almost exclusively concerns the raters' perspective. To obtain a comprehensive picture of CA, the test takers' self-assessments of CA and their similarities and differences with the raters' assessment also need to be investigated.

Acknowledgments

The publication of the work was supported by Guangzhou Panyu Polytechnic (grant number: 2023JG04). The author is grateful to the teachers and students at Information Engineering University in China, who generously participated in the study and to Dr. Fang Nan for her help with the data analysis.

8. References

Bosker, H. R., Pinget, A-F., Quené, H., Sanders, T. J. M., & 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

Bygate, M., Skehan, P., & Swain, M. (2013). *Researching pedagogic tasks: Second language learning, teaching, and testing*. London: Routledge. https://doi.org/10.4324/9781315838267

De Jong, J. H. A. L., & Van Ginkel, L. W. (1992). Dimensions in oral foreign language proficiency. In L. T. Verhoeven, & J. H. A. L. De Jong (Eds.), *The construct of language proficiency: Applications of psychological models to language assessment* (pp. 187-205). Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/z.62.19jon

De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237-254. https://doi.org/10.1080/15434303.2018.1477780

De Jong, N. H. (2023). Assessing second language speaking proficiency. *Annual Review of Linguistics*, *9*(1), 541-560. https://doi.org/10.1146/annurev-linguistics-030521-052114

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

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012b). 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). Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/lllt.32.06jon

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(4), 476-490. https://doi.org/10.1017/S026144480800551X

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, *11*(2), 125-144. https://doi.org/10.1177/026553229401100203

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

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

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379-399. https://doi.org/1177/0265532210364407

Higgs, T. V., & Clifford, R. (1982). The push toward communication. In T. V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher* (pp. 57-79). Lincolnwood, IL: National Textbook Company.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder, (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32) (pp. 1-20). Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/lllt.32.01hou

Hulstijn, J. H., Schoonen, R., De Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203-221. https://doi.org/10.1177/0265532211419826

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135-159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*(3), 475-505. https://doi.org/10.1017/S0272263112000150

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

Koizumi, R., & In'nami, Y. (2024). Predicting functional adequacy from complexity, accuracy, and fluency of second-language picture-prompted speaking, *System*, *120*, 103208. https://doi.org/10.1016/j.system.2023.103208

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145-164. https://doi.org/10.1016/j.system.2004.01.001

Kuiken, F., & Vedder, I. (2014a). Raters' decisions, rating procedures and rating scales. *Language Testing*, *31*(3), 279-284. https://doi.org/10.1177/0265532214526179

Kuiken, F., & Vedder, I. (2014b). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329-348. https://doi.org/10.1177/0265532214526174

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*(3), 321-336.

Kuiken, F., & Vedder, I. (2022a). Measurement of functional adequacy in different learning contexts: Rationale, key issues, and future perspectives. *TASK*, *2*(1), 8-32. https://doi.org/10.1177/0265532216663991

Kuiken, F., & Vedder, I. (2022b). 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). Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/rmal.3.14kui

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

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* (pp. 81-100). Amsterdam: European Second Language Association.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, *27*(4), 590-619. https://doi.org/10.1093/applin/aml029

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387-417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52-76. https://doi.org/10.1177/026553229000700105

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578. https://doi.org/10.1093/applin/amp044

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, *21*(1), 109-148. https://doi.org/10.1017/S0272263199001047

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

Pallotti, G. (2021). Measuring complexity, accuracy, and fluency (CAF). In P. Winke, & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 201-210). New York: Routledge. https://doi.org/10.4324/9781351034784-23

Pinget, A.-F., Bosker, H. R., Quené, H., & De Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, *31*(3), 349-365. https://doi.org/10.1177/0265532214526177

Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*, *36*(2), 161-179. https://doi.org/10.1177/0265532216669537

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *The Modern Language Journal*, *102*(4), 713-731. https://doi.org/10.1111/modl.12509

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, *33*(1), 53-73. https://doi.org/10.1177/0265532215579530

R Core Team. (2023). *R: A Language and Environment for Statistical Computing* [Software]. R Foundation for Statistical Computing. https://www.R-project.org/

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

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech? *TESOL Quarterly*, *50*(2), 421-446. https://doi.org/10.1002/tesq.234

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1-14. https://doi.org/10.1017/S026144480200188X

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532. https://doi.org/10.1093/applin/amp047

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*(3), 185-211. https://doi.org/10.1177/136216889700100302

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

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905-916. https://doi.org/10.1017/S1366728912000168

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 85-106. https://doi.org/10.1177/026553229701400105

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, *24*(1), 1-27. https://doi.org/10.1093/applin/24.1.1