BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

COMPLEXITY, ACCURACY AND FLUENCY (CAF) FEATURES OF SPEAKING PERFORMANCES ON APTIS ACROSS DIFFERENT LEVELS ON THE COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES (CEFR)

AR-G/2018/1

Xun Yan and Ha Ram Kim, University of Illinois at Urbana-Champaign
Ji Young Kim, University of California, Los Angeles

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITOR: VIVIEN BERRY
ISSN 2057-5203

ABSTRACT

This study investigates complexity, accuracy and fluency (CAF) features of speaking performances on the Aptis test across different Common European Framework of Reference for Languages (CEFR) levels, as an effort to examine criterion-related and cognitive validity evidence for the test. Benchmark speech sets from 125 examinees (25 sets from each level of A1-C) were sampled, each including responses to four speaking tasks, amounting to a total of 500 speech samples. An array of CAF features was measured, spanning six sub-components: lexical sophistication and appropriateness, grammatical complexity and accuracy, fluency, and pronunciation. These linguistic features were then subjected to both univariate and multivariate statistical analyses to identify distinguishing CAF features that can significantly predict examinees' CEFR levels.

The results of this study revealed distinguishing features in all three CAF components. Post-hoc comparisons showed significant differences on various features between all adjacent levels except for B2 and C. Findings of this study provide supporting evidence for the criterion-related and cognitive validity of the Aptis speaking test, evidencing the alignment between key criteria assessed in Aptis and components of speaking ability on the CEFR. The discriminating CAF features can also assist in rater calibration and training processes for the test.

Authors

Xun Yan is an assistant professor of Linguistics, Second Language Acquisition and Teacher Education (SLATE), and Educational Psychology at the University of Illinois at Urbana-Champaign, where he is also the supervisor of the English Placement Test (EPT). His current research interests include post-admission language assessments, scale development and validation, rater performance and training, assessment literacy for language teachers, and test score use in educational settings.

Ha Ram Kim is a PhD candidate in Linguistics and SLATE (Second Language Acquisition and Teacher Education) at the University of Illinois at Urbana-Champaign. She is the research assistant for the English Placement Test at the University of Illinois, managing test development and administration. Her research interests include language testing, research methodology, and second language acquisition, focusing on speaking and writing in a second or foreign language.

Ji Young Kim is an assistant professor in the Department of Spanish and Portuguese at the University of California, Los Angeles. Her research focuses on heritage language acquisition, bilingualism and experimental phonetics and phonology.

CONTENTS

1. RATIONALE AND AIMS 5

2. THEORETICAL FRAMEWORK AND BACKGROUND 5

2.1 Socio-cognitive framework for test validation 5
2.2 Complexity, accuracy and fluency in L2 speaking performance 5

3. METHODS 7

3.1 The Aptis spoken corpus 7
3.2 Performance features used in the study 8
3.2.1 Global judgments of speech 9
3.2.2 Fluency and pronunciation features 9
3.2.3 Complexity and accuracy features 10
3.3 Statistical analyses 10
3.3.1 Correlation-based analysis 10
3.3.2 Multivariate analysis 10

4. RESULTS 11

4.1 Descriptive statistics and correlation-based analyses 11
4.1.1 Global judgments of speech 11
4.1.2 Macro-level and micro-level fluency features 14
4.1.3 Lexico-grammatical complexity and accuracy 17
4.1.4 Pronunciation features 19
4.2 Multivariate analysis 23
4.2.1 Principal component analysis for global judgments 23
4.3.2 Principal component analysis for performance features 25

5. DISCUSSION 30

5.1 RQ 1: What CAF features in Aptis speaking performances are associated with different CEFR levels of speaking ability? 30
5.1.1 Global judgments 30
5.1.2 Fluency features 30
5.1.3 Lexico-grammatical complexity and accuracy features 31
5.1.4 Pronunciation features 31
5.2 RQ 2: Do test-takers across different CEFR levels display systematic differences on sub-components of CAF features on the Aptis speaking test? 32
5.2.1 Sub-components of CAF features represented on the Aptis speaking test 32
5.2.2 CAF features characterising the differences across CEFR levels 32

6. CONCLUSIONS AND IMPLICATIONS 34

REFERENCES 35

APPENDIX 1: Questionnaire for global judgment of speech 37

List of tables

Table 1: Aptis spoken corpus for analysis 7
Table 2: Demographic information of examinees in the Aptis spoken corpus 8
Table 3: Descriptive statistics for global judgment of speech (z-score) 11
Table 4: Descriptive statistics for macro- and micro-level fluency features 15
Table 5: Descriptive statistics for lexico-grammatical complexity features 17
Table 6: Descriptive statistics for pronunciation features (vowel duration difference) 20
Table 7: Factor loadings for global speech judgments 23
Table 8: Descriptive statistics of factor score for global judgments by CEFR level 24
Table 9: ANOVA results for global judgments and CEFR level 24
Table 10: Tukey post-hoc comparisons of global judgment factor scores 25
Table 11: Factor loadings for CAF features 26
Table 12: Descriptive statistics of factor score for CAF features by CEFR level 27
Table 13: ANOVA results for CAF dimensions and CEFR level 28
Table 14: Tukey post-hoc comparisons of CAF dimension 2 factor scores 29
Table 15: Tukey post-hoc comparisons of CAF dimension 1 factor scores 29
Table 16: CAF features characterising the differences across CEFR levels 33

List of figures

Figure 1: Boxplots for global judgments of speech by CEFR level 13
Figure 2: Boxplots for articulation rate, speech rate and mean length of run by CEFR level 16
Figure 3: Boxplots for lexico-grammar complexity features by CEFR level 18
Figure 4: Boxplots for normalised duration difference by CEFR level 21
Figure 5: Boxplots for normalised duration of long and short vowels by CEFR level 22
Figure 6: Boxplots for normalised duration of vowels in stressed and unstressed syllables by CEFR level 22
Figure 7: Scree plot for principal component analysis of global speech judgments 23
Figure 8: Boxplot of factor score for global judgments 24
Figure 9: Scree plot for principal component analysis of CAF features 26
Figure 10: Boxplot of factor score for CAF features 28

1. RATIONALE AND AIMS

This project investigates the complexity, accuracy and fluency (CAF) features of speaking performances on Aptis across different levels on the Common European Framework of Reference for Languages (CEFR). An array of CAF features was examined, which can be classified into six sub-components:

- lexical sophistication
- lexical appropriateness
- grammatical complexity
- grammatical accuracy
- fluency
- pronunciation.

Using a corpus-based approach, this study examined the macro and micro relationships amongst CAF features and CEFR performance levels, and identified representative CAF features that distinguish Aptis speaking performances across CEFR levels.

Findings of this study provide criterion-related and cognitive validity evidence for the Aptis speaking test, evidencing: (1) the alignment between key speaking criteria in Aptis and components of speaking ability on the CEFR; and (2) examinees' cognitive processes of speech production on Aptis. Discriminating CAF features can assist in rater calibration, certification, and scoring procedures.

2. THEORETICAL FRAMEWORK AND BACKGROUND

2.1 Socio-cognitive framework for test validation

This project is situated within the socio-cognitive framework for test development and validation (O'Sullivan & Weir, 2011; Weir, 2005). The validation framework contains five key components (i.e., cognitive validity, context validity, scoring validity, criterion-related validity and consequential validity) and stresses the interaction among different types of validity evidence in building a coherent validity argument for language tests. This project focuses on criterion-related and cognitive validity evidence for the Aptis speaking test, specifically, by investigating the CAF features of Aptis speaking performances.

2.2 Complexity, accuracy and fluency in L2 speaking performance

The construct of L2 performance and proficiency has long been recognised as multi-componential and multi-dimensional, comprising three principal components: complexity, accuracy, and fluency (Skehan, 1998; Ellis, 2008; Ellis & Barkhuizen, 2005). As such, CAF features have been widely used to characterise test performances and test-taker proficiency levels in both L2 speaking and writing assessments (Housen & Kuiken, 2009).

In speaking assessment, fluency can be measured by rate and pausing features at both macro and micro levels. While the macro approach often measures and correlates temporal features with other linguistic features or overall language proficiency, micro-level fluency features focus more on disfluencies (the cognitive processes of pausing and recovery). Validation research for L2 speaking assessments tends to examine macro-level fluency features for convenience of operationalisation, although examination of micro-level disfluency features can evidence interactions between fluency and speakers' automatic access to grammar and lexis (see, e.g., Clark & Fox Tree, 2002; Corley & Stewart, 2008; Dornyei & Kormos, 1998).

Complexity features of speaking performances have been mostly investigated in corpus linguistics, especially in studies of register differences between spoken and written discourse. Among the most influential is the work of Biber (1988) and his colleagues (e.g., Biber, Gray & Poonpon, 2011), who argue that complexity of spoken discourse is represented by clausal subordination, rather than abstract nouns and nominalisation at phrasal level. The proposed project follows this line of research by operationalising lexico-grammatical complexity of speaking performances in terms of clausal subordination.

Pronunciation, intelligibility and comprehensibility do not typically fall within the CAF framework. However, perceptions of comprehensibility and accentedness are closely associated with speaker's fluency (e.g., Derwing, Rossiter, Munro & Thomson, 2004; Trofimovich & Isaacs, 2012) and lexico-grammatical complexity features (e.g., Saito, Trofimovich & Isaacs, 2015). Therefore, we argue that pronunciation features should be included as a sub-component of accuracy for speaking assessment.

The Aptis speaking test is designed to measure communicative speaking ability, with four tasks targeting A1–C levels on the CEFR. An examination of the rating scales and the CEFR descriptors of speaking ability identifies five shared key criteria: pronunciation and intelligibility, fluency, lexical sophistication and appropriateness, grammatical complexity and accuracy, all falling within the CAF framework. However, very few studies have examined CAF features of Aptis speaking performances in relation to CEFR descriptors.

This project employs a corpus-based approach to examine: (1) the relationships amongst CAF features and holistic scores of speaking performance on the Aptis test; and (2) CAF features that characterise and distinguish speaking performances across different CEFR levels. Specifically, the project seeks to address the following research questions (RQs):

1. **What CAF features in Aptis speaking performances are associated with different CEFR levels of speaking ability?**
2. **Do test-takers across different CEFR levels display systematic differences on sub-components of CAF features on the Aptis speaking test?**

3. METHODS

This study employed a mixed-methods approach to examine CAF features that characterise and distinguish each performance level of the CEFR targeted in the Aptis speaking test, and the macro/micro relationships amongst CAF features and holistic scores of speaking performances on the Aptis speaking test. This section first introduces the Aptis spoken corpus and discusses the performance features (linguistic features and global judgments of speech) selected and analysed in this study. Next, how these performance variables were statistically analysed using both correlation-based and multivariate analyses are explained.

3.1 The Aptis spoken corpus

The spoken corpus used for this study comprised 125 sets of benchmark speech samples randomly drawn from responses to Tasks 3 and 4 on the Aptis speaking test. It included speech samples of 25 examinees from each level of A1–C (see Table 1 for details about each task). Each speech set consisted of four speech samples: three responses to Task 3 and one response to Task 4. This resulted in a total of 500 speech samples in the Aptis spoken corpus. All speech files were at first converted from mp3 to wav format, and then they were normalised in audacity for better sound quality and background noise reduction. Each speech sample was transcribed in a computer readable format using ELAN software by trained transcribers according to the PNC transcription guidelines.

The examinees were from around the globe, representing a wide range of first language (L1) backgrounds. As Table 2 shows, the five most represented countries were: India (25.6%), Saudi Arabia (10.4%), Colombia (8.8%), Mexico (8.8%) and Ukraine (7.2%). The five most presented L1 backgrounds were: Malayalam (24.8%), Spanish (20%), Arabic (14.4%), Ukrainian (7.2%), and Mandarin Chinese (4.8%). There were slightly more female examinees (52.8%) than male examinees (47.2%).

Table 1: *Aptis spoken corpus for analysis*

Task	Target CEFR level	Length	Skill focus	Number of examinees per level
Task 3	B1	135 secs (45 secs/ question)	Describing, comparing and contrasting, providing reasons and explanations to spoken questions	C: 25 B2: 25 B1: 25 A2: 25 A1: 25 Total: 125 exams
Task 4	B2	2 mins	Integrating ideas regarding an abstract topic into a long turn. Giving opinions, justifying opinions, giving advantages and disadvantages.	C: 25 B2: 25 B1: 25 A2: 25 A1: 25 Total: 125 exams

Table 2: *Demographic information of examinees in the Aptis spoken corpus*

Category	Level	N	%
L1 background	Malayalam	31	24.8%
	Spanish	25	20.0%
	Arabic	18	14.4%
	Ukrainian	9	7.2%
	Mandarin Chinese	6	4.8%
	Other	36	28.8%
Country	India	32	25.6%
	Saudi Arabia	13	10.4%
	Colombia	11	8.8%
	Mexico	11	8.8%
	Ukraine	9	7.2%
	Other	49	39.2%
Gender	Male	66	52.8%
	Female	59	47.2%
Total		**125**	**100%**

3.2 Performance features used in the study

Analyses of performance features in this study underwent three stages: (1) native-speaker global judgments of speech; (2) audio-based analysis of fluency and pronunciation features on Praat (Boersma & Weenink, 2015); and (3) transcript-based analysis of lexico-grammatical complexity and accuracy features through either manual coding or automated text evaluation tools. Detailed explanation below shows how performance features are analysed in each stage.

Participants' speech samples on the test were transcribed into a computer readable format by trained transcribers, following a consistent transcription convention (http://fave.ling.upenn.edu/downloads/Transcription_guidelines_FAAV.pdf). This transcription convention segments speech by silent pauses and thus allows for both acoustic analysis of pronunciation and fluency/disfluency features and transcript-based analysis of lexico-grammatical complexity and accuracy.

Both holistic and analytic approaches to operationalising CAF features were employed. The holistic CAF measures included native-speaker global judgments of each of the six criteria on a slider scale. The analytic CAF features were examined through acoustic and text analyses. Prior to acoustic analysis, all speech files were normalised to reduce background noise; this was done in order to generate more reliable results during the acoustic analysis. However, strong background noise remained in a number of speech files even after normalisation; therefore, these files were excluded from acoustic analysis of pronunciation features and some fluency/disfluency features. In addition, in transcription-based lexico-grammatical analyses, a number of speech files were also excluded because of the short text length. Therefore, in the descriptive statistics and correlation-based analyses, the sample sizes of individual performance features differed. However, in the multivariate analysis, we only included the files that do not have missing values; this resulted in a sample of 85 speech samples.

3.2.1 Global judgments of speech

The speech samples were evaluated on six criteria (intelligibility, comprehensibility, lexical sophistication, lexical appropriateness, grammar complexity and fluency) by five trained raters who were all graduate students in Linguistics or Speech and Hearing Sciences at the University of Illinois at Urbana-Champaign. Four of the raters were native-speakers of English and one rater was near-native speaker of English. After familiarising themselves with the Aptis tasks, they listened to each speech sample in a randomised order and holistically rated on each of the six criteria on a 5-point continuous sliding scale (e.g., Saito et al., 2015). See Appendix 1 for the questionnaire used for speech rating. Each speech sample was rated by two raters. Raters were overall consistent in their rating of the speech samples; rater reliability, as expressed in Spearman Rho correlation coefficients between the two raters, ranged between .55 to .78 on all six criteria ($r_{intelligibility}$ = .61; $r_{comprehensibility}$ = .72; $r_{lexical\_sophistication}$ =.78; $r_{lexical\_appropriateness}$=.55; $r_{grammatical\_complexity}$ = .74; $r_{fluency}$ = .70, $p < .001$).

However, for our analysis, to account for possible systematic score variance associated with rater severity, scores assigned by each rater was normalised to z-scores, and then the average of the two scores across the raters per each speaker was used.

3.2.2 Fluency and pronunciation features

Fluency features. The macro-level fluency features of each speech sample were automatically extracted using a Praat script developed by de Jong and Wempe (2009). These holistic temporal features include: (1) number of syllables, (2) speech rate, (3) articulation rate, (4) number of silent pauses, (5) mean length of run, and (6) number of fillers. Micro-level fluency/disfluency features were manually coded on: (1) pause type (i.e., silent and filled pause); (2) pause position (i.e., juncture and non-juncture pauses); (3) possible causes of pause (i.e., lexico-grammatical search and modification, and formulation of content); and (4) pause recovery (i.e., syllable lengthening, repeat, modification, and false start/restart). These micro-level disfluency features were further normalised and/or transformed into two variables for statistical analysis: (1) proportion of juncture pauses and (2) success in repairing non-juncture pauses.

Pronunciation features. Pronunciation features are important indices used in test validation (see, e.g., Anderson, Hsieh, Johnson & Koehler, 1992; Issaic & Trofimovich, 2012; Kang, Rubin & Pickering, 2010). However, on the Aptis speaking test, test-takers represent a wide range of backgrounds of English learning and use. Analysing the pronunciation accuracy of the Aptis test-takers is particularly challenging, because of the variability in examinees' L1 background and the varieties of English to which they are exposed. When multiple norms or target standards are represented among the examinees, it is difficult to determine whether L2 speakers correctly pronounce English sounds. Therefore, instead of focusing on deviation of particular sounds from the target norms, the present study examined sounds important to speech intelligibility. Specifically, we followed Jenkin's (2000) Lingua Franca Core (LFC) which focuses on phonetic features that are crucial to intelligibility among L2 speakers of English with various L1 backgrounds. These features include: intervocalic /t/ rather than flap [ɾ]; rhotic [ɻ] rather than other varieties of /r/; aspiration following the fortis plosives /p, t, k/; fortis/lenis differential effect on preceding vowel length; initial clusters not simplified; maintenance of vowel length contrasts and preservation of /ɜː/ (BIRD-type vowel).

The present study acoustically analysed these features using Praat (Boersma & Weenink, 2005). The Forced Alignment and Vowel Extraction (FAVE) program (Rosenfelder et al., 2011) was used to automatically align examinees' speech with the text transcription. This process allowed us to locate each phone in the speech signal without manually coding it.

3.2.3 Complexity and accuracy features

Complexity features. Lexical sophistication was automatically evaluated through Coh-Metrix (McNamara, Graesser, McCarthy & Cai, 2014) on lexical range (LMTD, vocd) and lexical frequency (CELEX log word frequency). Type-token ratio (TTR, Templin, 1957) was not used to measure lexical diversity as each speech sample had a wide range of length and TTR is sensitive to length (McCarthy & Jarvis, 2010). Instead, MLTD and vocd, which are modified TTR to adjust for text length, were used to represent lexical diversity. For grammatical complexity, number of subordination per c-unit was used.

Accuracy features. In terms of lexico-grammatical accuracy, errors were identified manually in the domains of syntax, morphology, and word order for grammatical accuracy. Lexical errors (i.e., incorrectly used or imprecise lexical expressions) were also identified for lexical appropriateness. Frequency of all errors was transformed into number of error-free clause per c-unit for analysis. In terms of analysing accuracy in lexico-grammar, speech samples in A1 were excluded, because most A1 speech samples comprised a list of words, not clauses. It is for the same reason the A1 level was excluded from the analysis of syntactic complexity.

3.3 Statistical analyses

Upon completion of global speech judgments and linguistic analyses, all performance variables were transformed into numeric variables; performance variables were averaged across the four tasks for all examinees and then screened for statistical analysis. The statistical analyses of this study included two phases: correlational analyses and multivariate analysis. Both phases of analyses were performed in R, using RStudio, version 1.1.383 (RStudio Team, 2016).

3.3.1 Correlation-based analysis

To address RQ 1, descriptive statistics of both global speech ratings and CAF features were examined for possible trends across score levels. Spearman rho correlation coefficients were computed among those CAF features and holistic speaking scores, as expressed in CEFR levels. These correlation coefficients were interpreted in light of which performance features tend to be associated with differences in speaking performance across CEFR levels.

3.3.2 Multivariate analysis

To answer RQ 2, all CAF features were incorporated in a two-step multivariate analysis to determine whether patterns of co-occurrence among CAF features can distinguish Aptis speaking performances across CEFR levels. First, to reduce Type I error in the significance tests, a principal component analysis with oblique rotation was performed to reduce the array of performance features to a smaller interpretable CAF dimensions. Oblique rotation was used because we assume that factors underlying CAF features are correlated, rather than orthogonal, to each other. Next, factor scores of the reduced dimensions were computed and subjected to univariate ANOVAs or MANOVAs to examine whether factor scores of CAF dimensions can distinguish speaking performance across CEFR levels.

4. RESULTS

4.1 Descriptive statistics and correlation-based analyses

4.1.1 Global judgments of speech

Table 3 shows descriptive statistics of global speech judgments across CEFR levels on the Aptis test. In general, a positive linear relationship with proficiency level was observed on all six measures (intelligibility, comprehensibility, lexical sophistication and appropriateness, grammatical complexity, and fluency), suggesting that higher proficiency group tend to produce speech that is more intelligible, comprehensible, lexically sophisticated and appropriate, grammatically complex, and fluent. Individual differences within each CEFR group could also be observed (see Figure 1); however, across the group, there tended to be a larger variance among the lower proficiency group (A1 and A2) but less variability among the higher proficiency group (B1, B2, and C), as revealed by large standard deviations.

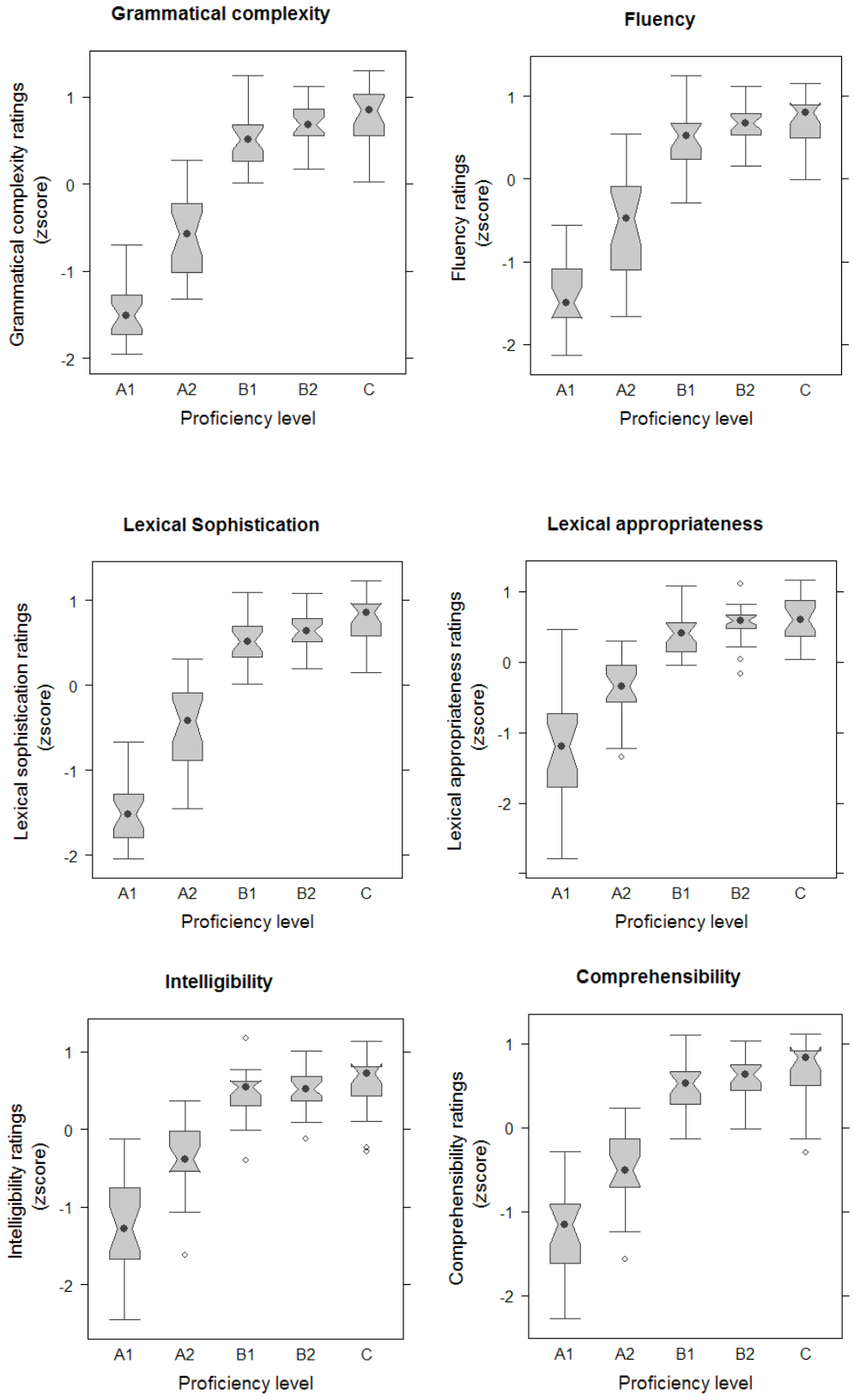
Spearman correlations were then run to examine the strength of associations between the mean ratings on these six criteria and proficiency level. As Table 3 shows, all six measures (intelligibility, comprehensibility, lexical sophistication and appropriateness, grammatical complexity, and fluency) strongly correlated with proficiency level indicated by CEFR level.

Table 3: *Descriptive statistics for global judgment of speech (z-score)*

Variable	CEFR level	N	Mean	SD	min	max
Intelligibility (*r* = .79)**	A1	25	-1.27	0.68	-2.45	-0.13
	A2	25	-0.36	0.46	-1.61	0.37
	B1	25	0.47	0.29	-0.39	1.17
	B2	25	0.51	0.28	-0.12	1.01
	C	25	0.6	0.35	-0.29	1.14
Comprehensibility (*r* = .82)**	A1	25	-1.27	0.56	-2.28	-0.28
	A2	25	-0.51	0.45	-1.57	0.23
	B1	25	0.49	0.3	-0.14	1.1
	B2	25	0.58	0.27	-0.01	1.04
	C	25	0.68	0.36	-0.29	1.12
Lexical Sophistication (*r* = .85)**	A1	25	-1.46	0.42	-2.04	-0.67
	A2	25	-0.5	0.52	-1.45	0.31
	B1	25	0.5	0.27	0.02	1.1
	B2	25	0.64	0.22	0.2	1.09
	C	25	0.77	0.31	0.15	1.23

Variable	CEFR level	N	Mean	SD	min	max
Lexical Appropriateness (*r* = .80)**	A1	25	-1.23	0.74	-2.79	0.46
	A2	25	-0.35	0.45	-1.35	0.3
	B1	25	0.38	0.27	-0.04	1.08
	B2	25	0.54	0.27	-0.17	1.11
	C	25	0.62	0.32	0.04	1.16
Grammatical Complexity (*r* = .85)**	A1	25	-1.44	0.38	-1.95	-0.69
	A2	25	-0.57	0.48	-1.32	0.27
	B1	25	0.51	0.3	0.02	1.25
	B2	25	0.67	0.23	0.17	1.12
	C	25	0.77	0.37	0.02	1.3
Fluency (*r* = .82)**	A1	25	-1.38	0.42	-2.13	-0.55
	A2	25	-0.53	0.58	-1.66	0.54
	B1	25	0.48	0.35	-0.29	1.25
	B2	25	0.66	0.26	0.16	1.12
	C	25	0.7	0.3	-0.01	1.15

Figure 1: *Boxplots for global judgments of speech by CEFR level*



4.1.2 Macro-level and micro-level fluency features

Table 4 presents descriptive statistics of macro- and micro-level fluency features among speaking performances across CEFR levels on the Aptis test. Overall, descriptive statistics for macro-level fluency variables showed clear increasing or decreasing trends that are aligned with findings in previous literature. In terms of rate features, speech rate, articulation rate and mean length of run all differed between lower and higher-proficiency test-takers. However, the correlation coefficients for speech rate and mean length of run were stronger than that for articulation rate (r_{ar} = .55, p < .01; r_{mlr} = .60, p < .01; r_{sr} = .70, p < .01), suggesting that speech rate and mean length of run are more closely associated with CEFR level and that these two variables are better proxies of fluency. This result makes sense in that the computation of both speech rate and mean length of run take into account silent pauses, while articulation rate omits silent pauses in speech. Moreover, boxplots for speech rate and mean length of run (see Figure 2) show that, whereas mean speech rate differed more at the lower end of the CEFR scale (A1 vs. A2 vs. B1), mean length of run differed more at the higher end (i.e., B1 vs. B2/C), although the variance at those levels for both variables were relatively large. In terms of pausing, number of silent pauses decreased with proficiency level (r_{sp} = -.59, p < .01). This suggests that higher proficiency examinees tend to produce more fluent runs and fewer long silent pauses.

As to micro-level disfluency features, test-takers at all CEFR levels paused frequently, in the form of either silent or filler pause. However, when silent pauses were further unpacked, higher proficiency examinees tended to pause more frequently at syntactic junctures (e.g., clausal boundaries). Pausing at these juncture positions is commonly observed in native-speaker speech, which is often regarded as expected pauses and can, to some extent, facilitate communication effectiveness by providing a cognitive break for the listener or interlocutor. In contrast, pausing at non-juncture positions is unexpected, tends to cause processing difficulty, and requires greater effort from the speaker to signal and repair the pauses. The correlation between the proportion of juncture pauses and CEFR level was very strong ($r_{juncture\_pause}$ = .84, p < .001), suggesting that the ability to parse and pause at the expected junctures is a very strong predictor of language proficiency; and in the Aptis spoken corpus, it was a stronger predictor than speech rate, mean length of run, and number of silent pauses. A closer qualitative analysis of speech transcripts suggests that these pauses tended to occur as a result of syntactic parsing or formulation of content. In contrast, non-juncture pauses occurred more often as a result of laboured search or retrieval of lexico-grammatical items than formulation of content. In addition, when non-juncture pauses occurred, higher proficiency examinees tended to be more successful at repairing them (r_{pause_repair} = .57, p < .01).

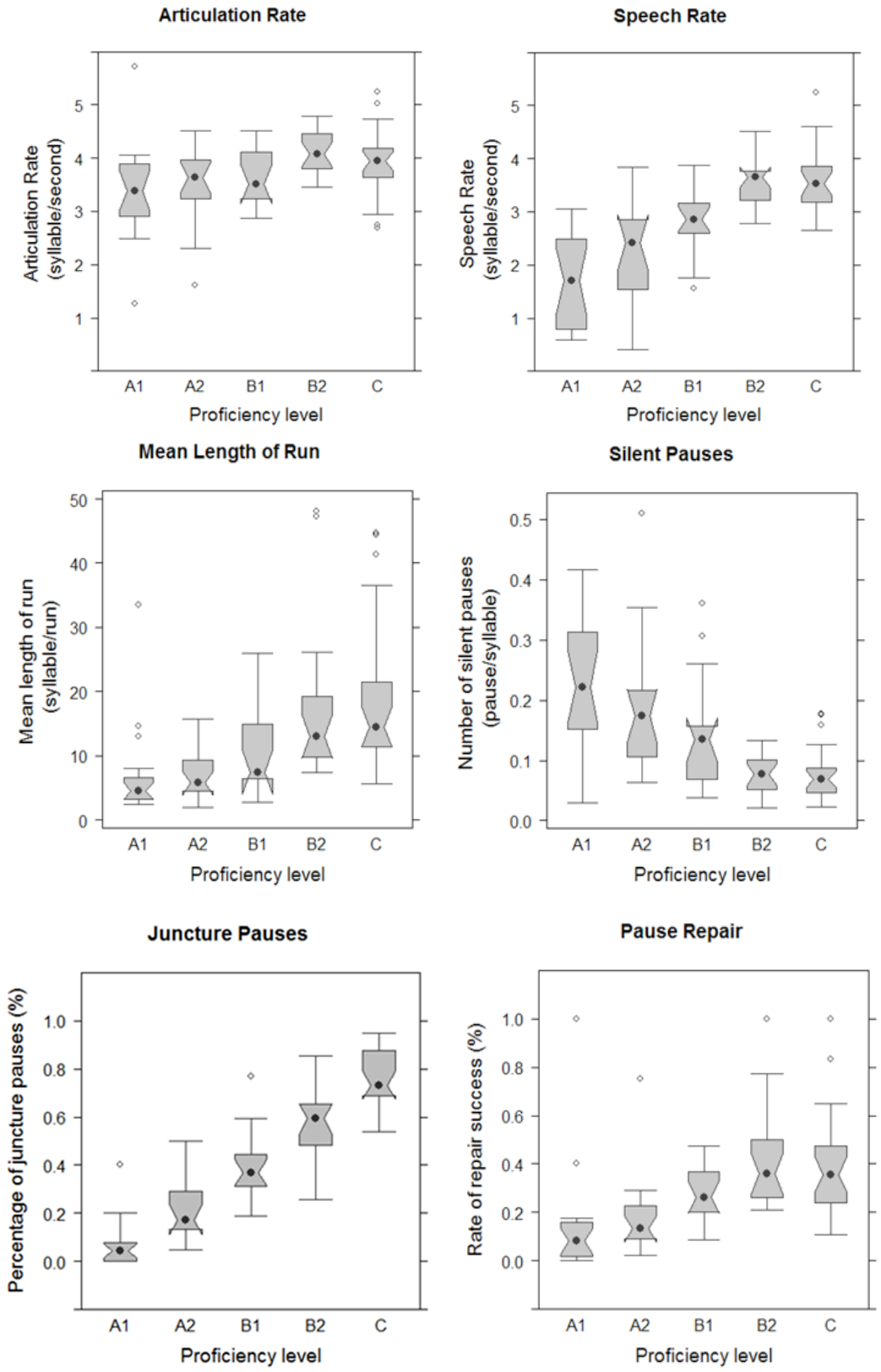
For filler pauses, interestingly, the raw number of pauses increased with proficiency. However, after normalising the fillers by number of syllables, the values were similar across CEFR levels (r_{filler_pause} = -.08, p = n.s.). This suggests that although higher-proficiency test-takers produced more fillers, it is simply because their responses were longer.

Taken together, the fluency and disfluency features appeared to distinguish examinees across multiple score levels (A1, A2, B1, B2/C); however, the differences in these variables are mostly negligible between B2 and C.

Table 4: *Descriptive statistics for macro- and micro-level fluency features*

Category	Variable	CEFR level	N	Mean	SD	min	max
Macro	Speech rate (r = .70**)	A1	18	1.7	0.83	0.59	3.05
		A2	16	2.28	0.99	0.41	3.84
		B1	15	2.79	0.67	1.55	3.88
		B2	20	3.54	0.42	2.79	4.52
		C	25	3.54	0.58	2.66	5.24
	Articulation rate (r = .55**)	A1	18	3.37	0.89	1.26	5.72
		A2	16	3.49	0.76	1.62	4.51
		B1	15	3.68	0.58	2.87	4.52
		B2	20	4.12	0.38	3.46	4.79
		C	25	3.91	0.63	2.69	5.24
	Number of silent pause (r = -.59**)	A1	18	0.23	0.11	0.03	0.42
		A2	16	0.19	0.12	0.06	0.51
		B1	15	0.14	0.1	0.04	0.36
		B2	20	0.08	0.03	0.02	0.13
		C	25	0.08	0.04	0.02	0.18
	Mean length of run (r = .60**)	A1	18	7.02	7.43	2.4	33.5
		A2	16	7.14	3.88	1.96	15.76
		B1	15	10.73	6.78	2.78	26
		B2	20	17.37	11.53	7.49	48.11
		C	25	18.57	11.66	5.66	44.75
Micro	Fillers (r = -.08)	A1	18	0.22	0.14	0.02	0.52
		A2	19	0.23	0.18	0.03	0.7
		B1	23	0.2	0.15	0.01	0.55
		B2	24	0.14	0.1	0.01	0.38
		C	21	0.21	0.13	0.05	0.5
	Juncture pauses (r = .84**)	A1	18	0.07	0.1	0	0.4
		A2	16	0.22	0.12	0.05	0.5
		B1	15	0.39	0.15	0.19	0.77
		B2	20	0.66	0.3	0.28	0.86
		C	25	0.75	0.35	0.43	0.93
	Pause repair (r = .57**)	A1	18	0.15	0.23	0	1
		A2	16	0.18	0.17	0.02	0.75
		B1	15	0.28	0.11	0.09	0.48
		B2	20	0.41	0.2	0.21	1
		C	25	0.41	0.25	0.11	1

Figure 2: *Boxplots for articulation rate, speech rate and mean length of run by CEFR level*



4.1.3 Lexico-grammatical complexity and accuracy

Complexity. Table 5 and Figure 3 show the descriptive statistics of complexity features across CEFR levels. A significant correlation with proficiency level was observed on lexical sophistication (LMTD and vocd), which suggested that higher proficiency speakers produced more sophisticated and diverse vocabulary. Between-group differences between A and B-level performances appeared to be larger than the differences between B and C-level performances. Specifically, differences in complexity variables between B2 and C were negligible. When it comes to grammatical complexity, similar trends were found. Proportion of subordinate clause per c-unit showed a linear trend aligned with proficiency level. While no difference was found between B2 and C, there seemed to be a distinguishable difference between A2 and B1, and B1 and B2/C performances. Spearman correlations showed strong associations between lexical sophistication and proficiency level (r_{lmtd} = .79, r_{vocd}= .79, $p < .01$), and moderate association between grammatical complexity and proficiency level ($r_{subordinate\_clause}$ = .51, $p < .01$); however, no associations were found with lexical frequency as indicated by CELEX log frequency. That is, there did not appear to be a meaningful difference across the proficiency level.

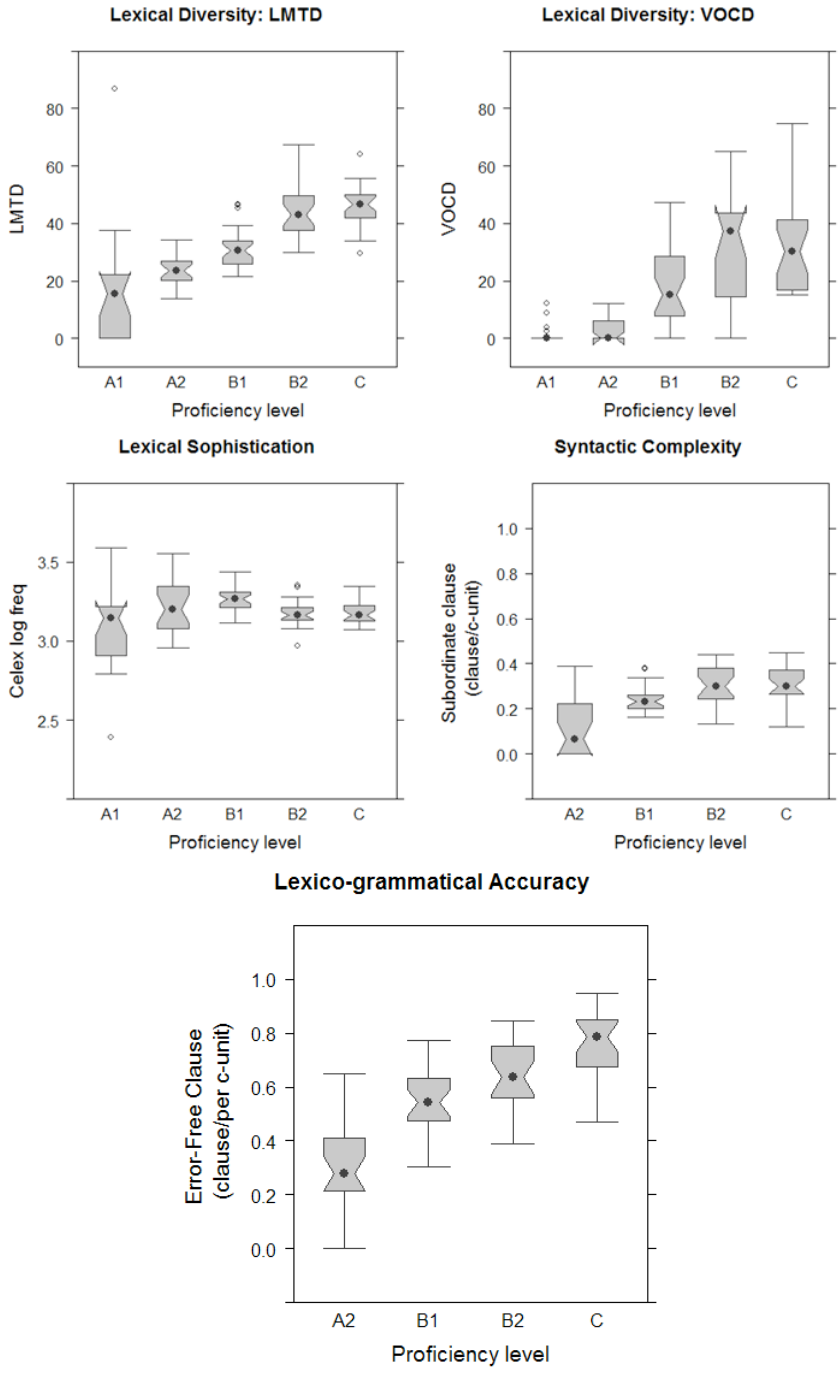
Accuracy. As Table 5 shows (the last four rows), proportion of error-free clause per c-unit linearly increased as proficiency level increased. Correlation coefficient also revealed a strong association between accuracy measure and proficiency level ($r_{error\_free\_clause}$ = .77, $p < .01$). The result suggested that higher-proficiency group produced lexically and grammatically more accurate speech in comparison to the lower-proficiency group.

Table 5: *Descriptive statistics for lexico-grammatical complexity features*

Variable	CEFR level	N	Mean	SD	min	max
LMTD (*r* = .79)**	A1	21	17.16	19.79	0	86.68
	A2	23	23.53	5.29	34.34	20.34
	B1	25	32.06	7.03	21.75	46.52
	B2	25	44.36	9.16	29.95	67.38
	C	25	45.63	7.75	29.52	63.93
VOCD (*r* = .79)**	A1	21	1.3	3.27	0	12.32
	A2	23	3.29	4.01	0	12.36
	B1	25	17.56	8.78	0	37.06
	B2	25	31.85	18.26	0	65.15
	C	25	34.21	19.22	15.33	74.89
CELEX log word frequency (*r* = .01)	A1	21	3.09	0.28	2.39	3.59
	A2	23	3.22	0.17	2.96	3.55
	B1	25	3.27	0.08	3.16	3.44
	B2	25	3.18	0.08	2.97	3.36
	C	25	3.18	0.07	3.07	3.35
Subordinate clause per c-unit (*r* = .51)**	A2	23	0.12	0.13	0	0.39
	B1	25	0.24	0.06	0.16	0.38
	B2	24	0.3	0.09	0.13	0.44
	C	25	0.3	0.09	0.12	0.45

Variable	CEFR level	N	Mean	SD	min	max
Error-free clause per c-unit (*r* = .77)**	A2	23	0.3	0.17	0	0.65
	B1	25	0.54	0.12	0.3	0.77
	B2	24	0.64	0.12	0.39	0.84
	C	25	0.76	0.13	0.47	0.95

Figure 3: *Boxplots for lexico-grammar complexity features by CEFR level*



4.1.4 Pronunciation features

Given the nature of the recording conditions of speaking tests (i.e., multiple people taking the test in the same room), there was a large amount of background noise in the present corpus. Consonants are typically vulnerable to ambient noise, because they are generally realised with low intensity as a result of constriction in the vocal tract (Cunningham et al., 2002; Stevens, 1998). Vowels, on the other hand, are more resistant to ambient noise, since they are produced with less constriction in the vowel tract and thus are realised with higher intensity, compared to consonants. Thus, the present study focuses on the production of vowels particularly the effects of vowel length (long vs. short vowels) and stress (stressed vs. unstressed) on vowel duration. The duration (in seconds) of 15 different types of vowels was extracted using the FAVE program (Rosenfelder et al., 2011), which were classified as either short vowels (i.e., BIT, BET, BAT, BOT, BUT, PUT) or long vowels (i.e., BEAT, BAIT, BOUGHT, BIRD, BOAT, BOOT, BITE, BOUT, BOY). With regard to stress, only vowels with primary stress and unstressed vowels were examined. In order to control for individual differences, the raw duration values were normalised using z-score normalisation. Speech files with unintelligible or no speech, severe background noise or errors in the FAVE program when extracting vowel properties were excluded from the analyses. In total, speech files of 18 participants (A1: 7, A2: 5, B1: 2, B2: 1, C: 3) were excluded in this process.

Table 6 and Figure 4 present descriptive statistics of the normalised duration difference between long and short vowels and the normalised duration difference between stressed and unstressed vowels across CEFR levels. As vowel length and stress are closely related each other (e.g., long vowels in stressed syllables tend be realised with longer duration than those in unstressed syllables), the dataset was divided into four variables: (1) duration difference between long and short vowels in stressed syllables; (2) duration difference between long and short vowels in unstressed syllables; (3) duration difference between stressed and unstressed vowels for long vowels; and (4) duration difference between stressed and unstressed vowels for short vowels.

Overall, the mean duration difference values were similar across the CEFR levels. Spearman correlation coefficients for the four variables were weak and statistically non-significant. However, averaging the duration differences may be problematic, as it obscures valuable information of the individual tokens. Thus, we performed mixed-effects models on the normalised duration of long vs. short vowels and on stressed vs. unstressed vowels, to see if the analysis of repeated measures revealed any difference.

Table 6: *Descriptive statistics for pronunciation features (vowel duration difference)*

Variable	CEFR level	N	Mean	SD	min	max
Duration difference (normalised) between long vs. short vowels stressed syllables) (*r* = .15)	A1	18	0.33	0.55	-0.4	1.5
	A2	20	0.26	0.34	-0.37	0.99
	B1	23	0.31	0.18	-0.07	0.65
	B2	24	0.38	0.17	0.11	0.67
	C	22	0.38	0.14	0.08	0.62
Duration difference (normalised) between long vs. short vowels (unstressed syllables) (*r* = .05)	A1	18	0.38	0.71	-1.44	1.09
	A2	20	0.47	0.49	-0.36	1.55
	B1	23	0.47	0.32	-0.01	1.05
	B2	24	0.56	0.37	-0.63	1.05
	C	22	0.56	0.29	-0.25	1.05
Duration difference (normalised) between stressed vs. unstressed vowels (long vowels) (*r* = .07)	A1	18	0.28	0.66	-0.83	1.41
	A2	20	-0.15	0.52	-1.64	0.71
	B1	23	0.03	0.24	-0.33	0.67
	B2	24	0.09	0.26	-0.35	0.63
	C	22	0.11	0.18	-0.28	0.46
Duration difference (normalised) between stressed vs. unstressed vowels (short vowels) (*r* = .03)	A1	18	0.34	0.51	-0.76	1
	A2	20	0.06	0.38	-0.6	0.92
	B1	23	0.2	0.24	-0.34	0.56
	B2	24	0.27	0.3	-0.26	0.79
	C	22	0.29	0.31	-0.48	0.82

Figure 4: *Boxplots for normalised duration difference by CEFR level*

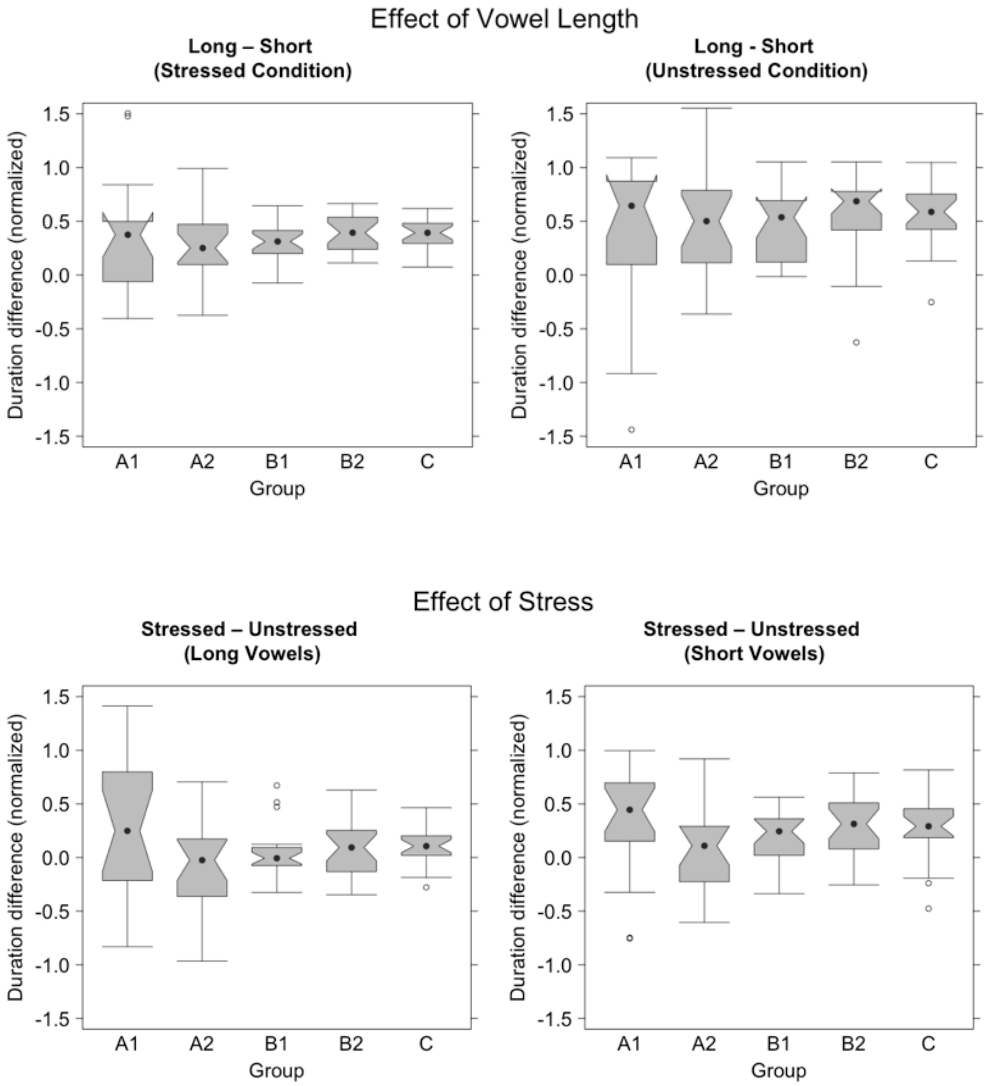


Figure 5 shows the normalised duration of long and short vowels in stressed and unstressed syllables. The effects of group and vowel length and the interaction between the fixed factors on the duration of stressed and unstressed vowels were analysed using linear mixed effects modeling with examinees and items as random effects. The *lmer* function in the *lme4* package (Bates et al., 2015) in R was used for the analyses and further pairwise analyses were conducted using the *lsmeans* function in the *lsmeans* package (Lenth, 2016). Results showed that there was a main effect of vowel length on the duration of vowels in both stressed and unstressed syllables (stressed: β = -0.379, SE = 0.138, t = -2.741, $p < .05$; unstressed: β = -0.515, SE = 0.146, t = -3.532, $p < .01$). With regard to the vowels in stressed syllables, a marginally significant interaction between group (C-level) and vowel length was found (β = -0.193, SE = 0.09, t = -2.141, p = 0.062). That is, C-level examinees' duration difference between long and short vowels was larger compared to that of the A1-level examinees (baseline group).

Figure 5: *Boxplots for normalised duration of long and short vowels by CEFR level (left: stressed syllable, right: unstressed syllable)*

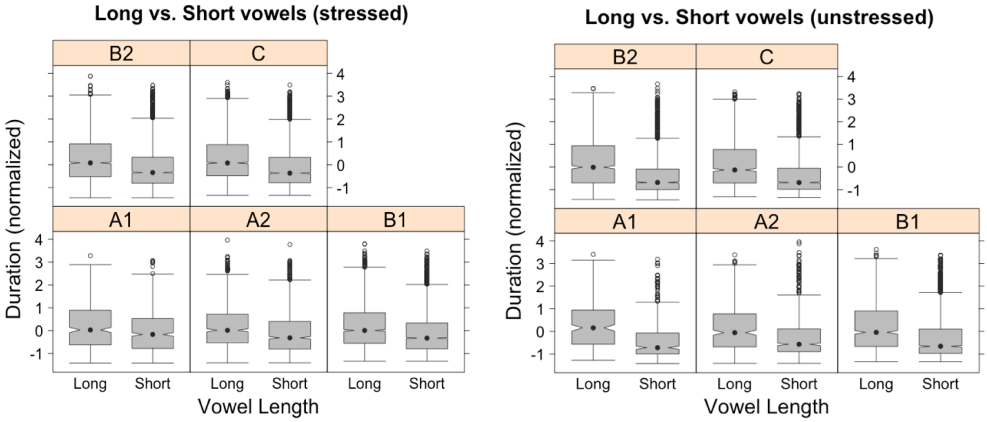
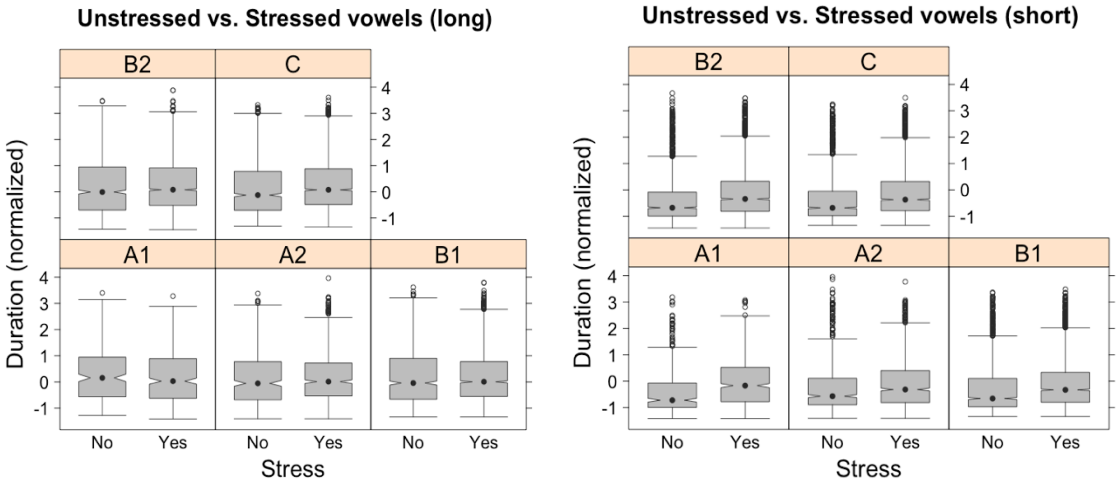


Figure 6 shows the normalised duration of vowels in stressed and unstressed syllables. The graph on the left demonstrates the duration of long vowels and the graph on the right demonstrates the duration of short vowels. The effects of group and stress and the interaction between the fixed factors on the duration of long and short vowels were analysed using linear mixed effects modeling with examinees and items as random effects. Results showed that there was a main effect of stress on the duration of long and short vowels (long: β = -0.13, SE = 0.03, t = -4.05, p < .001; short: β = 0.27, SE = 0.02, t = 13.50, p < .001). For long vowels, a marginally significant interaction was found between group (C-level) and stress (β = 0.19, SE = 0.10, t = 1.92, p = .05), indicating that C-level examinees' duration difference between stressed and unstressed vowels was larger than that of the A1-level examinees. Significant interactions between group (A2-, B1-, B2-levels) and stress were also found in short vowels (A2 * stress: β = -0.20, SE = 0.06, t = -3.01, p < .01; B1 * stress: β = -0.18, SE = 0.06, t = -2.94, p < .01; B2 * stress β = -0.12, SE = 0.06, t = -1.97, p < .05). That is, when producing short vowels, the A1-level examinees distinguished stressed and unstressed vowels to a larger degree than A2-, B1- and B2-level examinees. C1-level examinees' duration difference between stressed and unstressed vowels did not differ from that of the A1-level examinees.

Figure 6: *Boxplots for normalised duration of vowels in stressed and unstressed syllables by CEFR level (left: long vowels, right: short vowels)*



4.2 Multivariate analysis

Principal component analysis was performed on global judgments by human raters and linguistic features separately, as both sets of features cover the same CAF domains. Moreover, performing separate PCA on these features allows for comparison as to what extent each set of features can predict examinees' CEFR levels.

Prior to running the PCA, all the performance feature data were screened for required statistical assumptions of PCA. The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy for all performance features was .88, which is above the suggested minimum of .6 for principal component analysis. The skewness and kurtosis of all features are within the range of (-3, 3), suggesting that all the features are approximately normally distributed. Correlation coefficients among all features are below .9, except for that between LTMD and vocd. Since the two features measure the same construct (i.e., lexical diversity), we excluded vocd from the PCA. These procedures resulted in a set of six global judgment features and 15 CAF features. Taken together, all these statistics suggest that the data were adequate for principal component analysis.

4.2.1 Principal component analysis for global judgments

The scree plot of the PCA for global judgments (see Figure 7) suggests a single factor solution. The factor loadings for the rating features (see Table 7) were all strong, suggesting that human raters assigned similar ratings across the six features. Overall, this factor accounted for 94.93% of the variance and covariance among the global ratings.

Figure 7: *Scree plot for principal component analysis of global speech judgments*

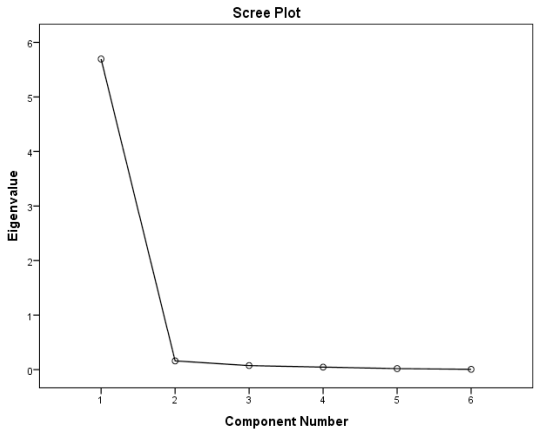


Table 7: *Factor loadings for global speech judgments*

Variable	Factor loading
Grammatical complexity	.98
Lexical sophistication	.98
Comprehensibility	.97
Intelligibility	.97
Fluency	.96
Lexical appropriateness	.96

Table 8 and Figure 8 present the descriptive statistics and boxplot for the factor score of global judgments across CEFR levels. There was a strong correlation between factor score and CEFR ($r_{global\_judgment\_factor\_score}$ = .80, p < .01). Table 9 shows the ANOVA results for the examinees' factor scores of global judgments and CEFR levels. There was a significant difference in the factor score of global speech judgments across CEFR levels. Tukey post-hoc comparisons, as shown in Table 10, suggest that global speech judgments could statistically distinguish three CEFR level groups: A1 < A2 < B1/B2/C; on this dimension, A2-level examinees performed better than did A1-level examinees; while examinees at and above B1 levels outperformed examinees at A1 and A2 levels, there was not significant difference among B1, B2 and C levels.

Table 8: *Descriptive statistics of factor score for global judgments by CEFR level*

CEFR level	N	Mean	SD	Min	Max
A1	16	-1.63	0.60	-2.69	-0.53
A2	15	-0.56	0.49	-1.54	0.11
B1	14	0.46	0.38	-0.18	1.33
B2	20	0.63	0.30	0.04	1.23
C	21	0.73	0.41	-0.20	1.29

Figure 8: *Boxplot of factor score for global judgments*

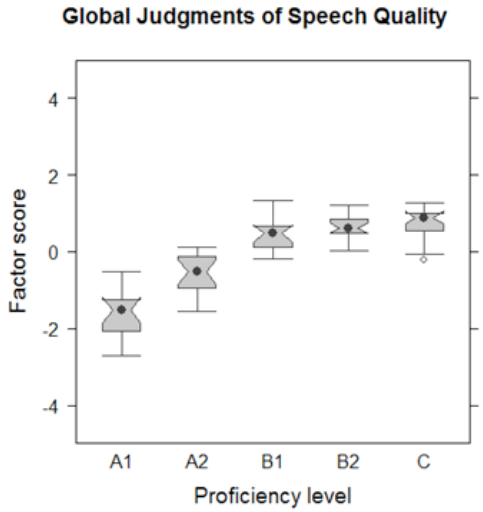


Table 9: *ANOVA results for global judgments and CEFR level*

	Sum of squares	*df*	Mean square	*F*	*p*
Between groups	69.21	4	17.30	88.75	<.001
Within groups	15.79	81	0.20		
Total	85.00	85			

Table 10: *Tukey post-hoc comparisons of global judgment factor scores*

CEFR level	N	Subset[†]		
		1	2	3
A1	16	-1.63		
A2	15		-0.56	
B1	14			0.46
B2	20			0.63
C	21			0.73

[†] Each subset column represents a significantly distinguishable group of CEFR levels; the levels within each subset column are not significantly different from one another.

4.3.2 Principal component analysis for performance features

The scree plot of the PCA for CAF features (see Figure 9) suggests a four-factor solution. Table 11 shows the factor loadings for the CAF features. Based on the factor loadings, the four components can be interpreted as follows:

- Component 1: Automaticity of lexico-grammar use
- Component 2: Macro-level fluency features
- Component 3: Pronunciation-short vowels
- Component 4: Pronunciation-long vowels

Interestingly, Component 1, the component that accounted for most variance/covariance among the CAF features (45.16%), loaded on both micro-level pausing features and lexico-grammatical complexity and accuracy. This dimension can be interpreted as the automaticity of lexico-grammar use. As discussed above, the occurrence of non-juncture pauses can suggest laboured search of lexico-grammatical items, an indicator of examinees' automaticity of lexico-grammatical knowledge. Thus, it makes sense that these features co-occur with lexico-grammatical complexity and accuracy features.

In terms of pronunciation features, two dimensions were extracted, although these two dimensions accounted for the least amount of variance/covariance (10.95% and 8.76% respectively). As is shown in Table 11 (the last four rows), the two pronunciation dimensions can be interpreted together, in that the first dimension is related to production of vowels or syllables with shorter duration (as a result of lexical stress), while the second dimension relates to the production of those with longer duration. Taken together, these four factors accounted for 76.24% of the variance and covariance among the CAF features.

Figure 9: *Scree plot for principal component analysis of CAF features*

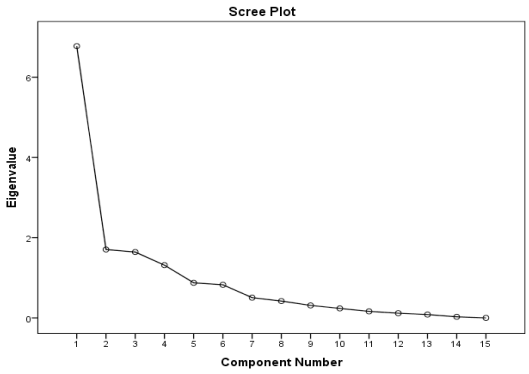


Table 11: *Factor loadings for CAF features*

Variable	Loadings			
	Component 1	Component 2	Component 3	Component 4
Pause repair	.94			
Juncture pauses	.86			
Subordinate clause	.75			
Mean length of run	.75			
Error-free clause	.72			
LMTD	.44			
Articulation rate		.97		
Speech rate		.86		
Number of syllables		.79		
Silent pauses		-.67		
Word count	.43	.56		
Length (unstressed syllables)			.89	
Stress (short vowels)			.82	
Stress (long vowels)				.94
Length (stressed syllables)				.77

Table 12 and Figure 10 present the descriptive statistics and boxplot for the factor score of global judgments across CEFR levels. The correlations between the first two dimensions and CEFR level were moderately strong to strong ($r_{automaticity\_factor\_score}$ = .87, $p < .01$; $r_{fluency\_factor\_score}$ = .66, $p < .01$), while the correlations for the pronunciation dimensions were not statistically significant. As discussed above, the lack of significant correlation for pronunciation dimensions might be due to the nature of the recording and the resultant narrow range of pronunciation features used for analysis.

A one-way MANOVA of the dimension scores among the Aptis speaking performances revealed significant multivariate main effect for score level (Wilks' $\lambda = .16$, $F(16,238.93) = 12.51$, $p < .001$, $\eta^2 = .37$). Tables 13 to 15 show the post-hoc ANOVA results for the examinees' factor scores of CAF dimensions and CEFR levels. There were significant differences across CEFR levels in only two dimensions: (1) automaticity of lexico-grammar use; and (2) macro-level fluency. Tukey post-hoc comparisons suggest that these two dimensions distinguish speaking performances at different adjacent CEFR levels. Specifically, on the macro-level fluency dimension (see Table 14), examinees at and above B1 levels outperformed examinees at A1 and A2 levels, suggesting that B/C-level examinees produce faster and longer speech with fewer pauses. In addition, B1-level examinee performed worse than did C-level examinees. However, there was no significant difference between A1 and A2, B1 and B2, or B2 and C.

In contrast, automaticity of lexico-grammar use statistically distinguished four CEFR level groups: A1 < A2 < B1 < B2/C; on this dimension (see Table 15), B2/C-level examinees outperformed B1-level examinees, followed by A2 and A1 levels. This suggests, as the proficiency level increases, examinees are more automatic at retrieving complex lexico-grammatical resources and maintaining an acceptable level of accuracy or appropriateness in use. However, there was not significant difference between B2 and C levels.

Table 12: *Descriptive statistics of factor score for CAF features by CEFR level*

Component	CEFR level	N	Mean	SD	Min	Max
Component 1 (*r* = .87)**	A1	16	-1.39	0.35	-1.93	-0.45
	A2	15	-0.74	0.46	-1.42	0.65
	B1	14	0.02	0.24	-0.40	0.45
	B2	20	0.70	0.58	-0.06	2.35
	C	21	0.91	0.57	-0.03	2.01
Component 2 (*r* = .66)**	A1	16	-1.04	0.81	-2.42	0.18
	A2	15	-0.67	1.16	-3.09	0.88
	B1	14	0.03	0.70	-1.39	1.30
	B2	20	0.71	0.39	0.12	1.70
	C	21	0.58	0.54	-0.53	1.93
Component 3 (*r* = .07)	A1	16	-0.31	1.47	-4.18	1.41
	A2	15	-0.11	0.96	-1.82	1.54
	B1	14	0.01	0.83	-1.38	1.31
	B2	20	0.32	0.73	-1.13	1.51
	C	21	0.00	0.91	-2.47	1.32
Component 4 (*r* = .16)	A1	16	-0.39	1.75	-1.76	4.04
	A2	15	-0.33	1.29	-3.95	1.28
	B1	14	-0.19	0.40	-0.97	0.65
	B2	20	0.05	0.50	-0.85	0.97
	C	21	0.11	0.47	-0.87	1.13

Figure 10: *Boxplot of factor score for CAF features*

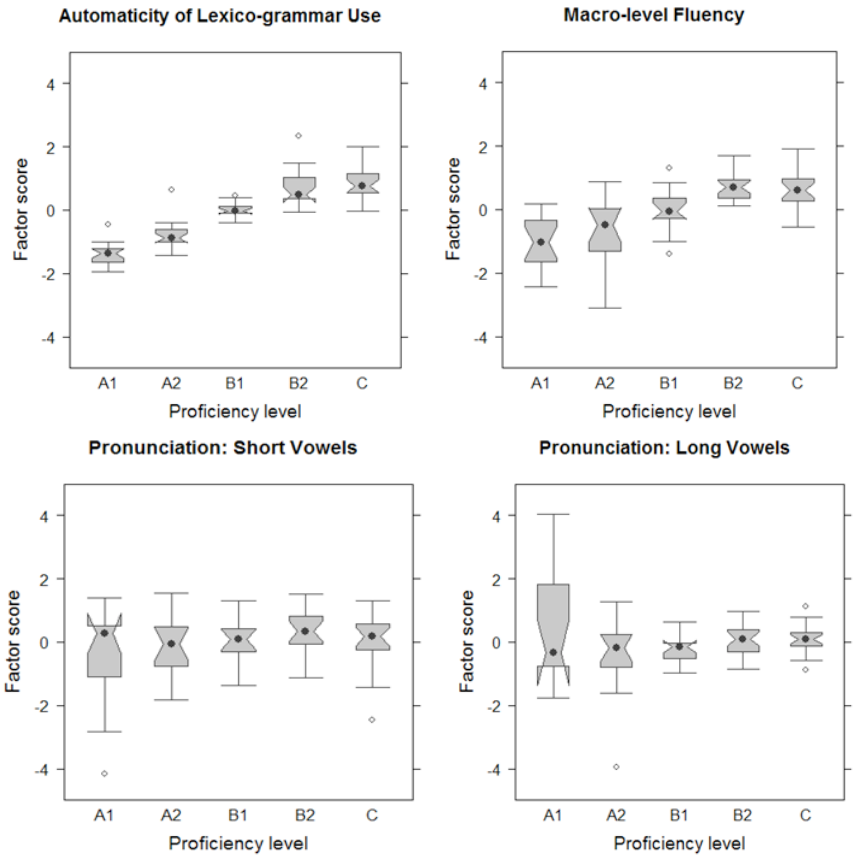


Table 13: *ANOVA results for CAF dimensions and CEFR level*

Dimension		Sum of squares	*df*	Mean square	F	*p*
Component 1	Between Groups	66.35	4	16.58	72.03	<.001
	Within Groups	18.65	81	0.23		
	Total	85.00	85			
Component 2	Between Groups	41.02	4	10.25	18.88	<.001
	Within Groups	43.98	81	0.54		
	Total	85.00	85			
Component 3	Between Groups	3.74	4	0.93	0.93	.45
	Within Groups	81.26	81	1.00		
	Total	85.99	85			
Component 4	Between Groups	4.82	4	1.20	1.21	.31
	Within Groups	80.18	81	0.99		
	Total	85.00	85			

Table 14: *Tukey post-hoc comparisons of CAF dimension 2 factor scores*

CEFR level	N	Subset†		
		1	2	3
A1	16	-1.17		
A2	15	-.69		
B1	14		.03	
B2	20		.62	.62
C	21			.73

† Each subset column represents a significantly distinguishable group of CEFR levels; the levels within each subset column are not significantly different from one another.

Table 15: *Tukey post-hoc comparisons of CAF dimension 1 factor scores*

CEFR level	N	Subset†			
		1	2	3	4
A1	15	-1.36			
A2	15		-0.75		
B1	14			-.015	
B2	20				0.66
C	21				0.89

† Each subset column represents a significantly distinguishable group of CEFR levels; the levels within each subset column are not significantly different from one another.

5. DISCUSSION

5.1 RQ 1: What CAF features in Aptis speaking performances are associated with different CEFR levels of speaking ability?

5.1.1 Global judgments

In this study, quality of speech in the Aptis spoken corpus was evaluated holistically by human raters with respect to six scoring criteria: intelligibility, comprehensibility, fluency, grammatical complexity, lexical sophistication, lexical appropriateness. Ratings on all six criteria were strongly related to CEFR levels, suggesting that there are meaningful increasing trends for all target speaking-related constructs with proficiency level. These strong correlations provide supporting evidence for the criterion-related validity of the Aptis speaking test.

That said, the global ratings were highly correlated with each other. After z-transformation, correlation coefficients among the ratings on the six criteria were all above .9. The rather strong correlations among the global judgment ratings can be ascribed to three possible explanations: (1) these features are highly correlated in nature; (2) human raters had halo effects when rating these features; and (3) differences in the original ratings across criteria were masked during the process of z-transformation. From findings of previous literature, it is evident the first explanation is unlikely, and that there are differences in descriptive statistics across different scoring criteria. A close examination of the original speech judgments reveals a good amount of variation in the raw ratings across criteria on individual examinees. Although this cannot completely rule out the possibility of a real halo effect among the raters, at the very least, the raters did not assign the same or very similar scores across the board, suggesting some level of independence among the global judgments. Therefore, the strong correlations among the speech ratings are more likely a result of z-score transformation. However, because z-scores only reflect relative standings of examinees rather than original score assignment on each criterion, we recommend using factor score of all criteria as a composite score of global speech quality (see Figure 8), when interpreting the relationship between speech quality ratings and CEFR levels.

5.1.2 Fluency features

We examined both macro-level and micro-level fluency features. Macro-level fluency features included widely used temporal features that measure the amount, rate, and pausing dimensions of speech fluency. Results of this study align with findings of previous fluency literature in that test-takers of higher proficiency tend to produce longer speech, speak with higher speech rate, and make fewer silent pauses. In terms of magnitude, these features showed moderately strong to strong correlations with CEFR levels, which is also similar with findings in previous fluency literature. In contrast, filler pauses did not show a meaningful relationship with proficiency level. The lack of association between filler pauses and CEFR levels might be due to two reasons. First, filler pauses can occur regardless of language proficiency level. Filler pauses, be it lexical (e.g., you know) or non-lexical (e.g., uh), tend to be a common phenomenon in speech production; they do not necessarily occur as a result of low proficiency or impede comprehension or interaction. In fact, in certain contexts, filler pauses can facilitate communication between interlocutors. This phenomenon has been well documented in fluency studies of L1 speech (e.g., Clark & Fox Tree, 2002), where filler such as *uh* and *um* might carry specific functions in communication and signal speech planning and break for the interlocutors. That is, the use of filler pauses can enhance communication effectiveness, because they provide a

cognitive break for the listeners in comprehension. However, these functions were not consistently observed in L2 speech (e.g., Bosker, Quené, Sanders & de Jong, 2014). Second, even when filler pauses occur because of a low level of language proficiency, the presence or amount of filler pauses alone might not mark language proficiency; rather, it is the recovery of pauses, or how filler pauses are handled, that is at stake. Therefore, pauses should be examined in conjunction with other micro-level disfluency features, e.g., pause position, and pause recovery.

In this study, micro-level fluency features were transformed into two variables: proportion of juncture pauses, and success rate of pause repair. Both variables were strongly correlated with proficiency level. That is, while pauses occur frequently in speech across all CEFR levels, as proficiency increases, there is a higher percentage of pauses that occur at syntactic junctures. This feature showed the strongest correlation with CEFR level among all features in this study. Non-juncture pauses occur mostly as a result of laboured search of lexico-grammar. Among these unexpected pauses, B2- and C-level test-takers are more capable of repairing those pauses. Taken together, the results of micro-level fluency features suggest that fluency (on the Aptis speaking test) is not merely about amount and rate of speech production, but also about automaticity in lexico-grammar use.

When these features are interpreted in relation to lexico-grammar, fluency is no longer a simple collection of holistic temporal features, but rather a representation of the cognitive process during speech production. Therefore, these fluency features can provide supporting evidence for the cognitive validity of the Aptis speaking test.

5.1.3 Lexico-grammatical complexity and accuracy features

Lexico-grammatical features in this study covered lexical diversity, lexical sophistication, syntactic complexity, and grammatical accuracy. All features showed moderate to strong positive correlations with CEFR levels except for lexical sophistication, suggesting that as proficiency level increases, the speech produced by test-takers becomes more lexico-grammatically complex and accurate. In addition, the correlation for lexical diversity was stronger than that for syntactic complexity. These results are in line with previous CAF studies in both SLA and language testing.

However, contrary to many previous CAF studies, lexical sophistication, operationalised as CELEX log-transformed vocabulary frequency, did not show a significant correlation with language proficiency. This unexpected result might be because of the nature of the task. The tasks on the Aptis speaking test require test-takers to describe a set of pictures and a personal experience in relation to the context or topic of the pictures. This kind of picture description tasks tends to elicit discourse features that are quite different from argumentative speaking that ask test-takers to summarise, compare and contrast, or argue for a position on a particular topic. However, further evidence is needed in order to falsify this hypothesis.

5.1.4 Pronunciation features

In terms of pronunciation features, we employed generalised linear mixed-effect model to examine the production of vowels, particularly the effects of vowel length (long vs. short vowels) and stress (in stressed vs. unstressed syllable) on vowel duration. The results showed significant main effects for both vowel length and stress, suggesting that overall, test-takers on the Aptis speaking test tend to be able to distinguish long and short vowels and syllables with and without lexical stress. In terms of proficiency level, no statistically significant difference was found across proficiency level in terms of test-takers' ability to distinguish vowel length and lexical stress. However, there was a trend that C-level test-takers tend to perform better at distinguishing: (1) long vs. short vowels, and (2) stressed vs. unstressed long vowels than did A-1 level test-takers.

When the repeated measures of all vowel productions were averaged across four conditions, namely, long unstressed, long stressed, short unstressed and short stressed vowels, the differences disappeared. This is probably because the duration differences across the 15 types of vowels are rather large that averaging across them masked the differences of vowel production among test-takers across proficiency levels. Given the relatively small size of the Aptis spoken corpus, we used the averaged duration measures for the multivariate analysis, to reduce Type I error rate. However, if a sufficiently large spoken corpus is available, it will be desirable to use the duration measures for each type of vowel in the analysis.

Taken together, the majority of the performance features examined in this study show expected relationships with CEFR levels. Following the socio-cognitive framework of test validation (O'Sullivan & Weir, 2011; Weir, 2005), these relationships provide supporting evidence for both criterion-related and cognitive validity of the Aptis speaking test.

5.2 RQ 2: Do test-takers across different CEFR levels display systematic differences on sub-components of CAF features on the Aptis speaking test?

5.2.1 Sub-components of CAF features represented on the Aptis speaking test

In order to reduce Type I error rate, we performed PCA on both the global judgments of speech quality and CAF features. The wide array of performance features was reduced to six dimensions: (1) global judgments of speech quality; (2) macro-level fluency; (3) automaticity of lexico-grammar use; (4) filler pauses; (5) pronunciation-short vowels; and (6) pronunciation-long vowels. These dimensions are all construct-relevant as they correspond to the rating criteria of the Aptis speaking test. Additionally, the features associated with each dimension are similar with findings of previous factor analytic studies of CAF features. These findings provide supporting evidence for the construct validity of the Aptis speaking test.

Interestingly, automaticity of lexico-grammar use loaded on lexico-grammatical complexity, accuracy, and micro-level fluency features. The macro-level fluency features clustered as a different factor (i.e., macro-level fluency). In contrast, micro-level fluency features clustered with lexico-grammatical features. The emergence of this factor aligns with earlier discussion that fluency should not simply be viewed as temporal features; the interpretation of fluency should include another dimension: fluency as a cognitive process, in particular, fluency as automaticity in language use. These two dimensions, though both crucial to speech fluency, are associated with different aspects of language proficiency.

5.2.2 CAF features characterising the differences across CEFR levels

Among the six subcomponents, only three components showed significant differences across CEFR levels: (1) global judgments of speech quality; (2) macro-level fluency; and (3) automaticity of lexico-grammar use. However, these three dimensions characterise systematic linguistic differences at different adjacent levels. Table 16 summarises results of post-hoc comparisons on all dimensions at each pair of adjacent levels on the Aptis speaking test. According to Table 16, automaticity of lexico-grammar use distinguished four of the five CEFR levels (see the second row). Both macro-level fluency and global judgments distinguished three CEFR levels (see the third and fourth rows); however, they distinguished different CEFR levels. Each column of Table 14 shows the difference(s) between adjacent CEFR levels. Specifically, A1 and A2 levels were significantly different on

automaticity of lexico-grammar use and global judgment of speech quality. A2 and B1 levels were significantly different on automaticity of lexico-grammar use, global judgment of speech quality, and macro-level fluency. B1 and B2 levels were only different on automaticity of lexico-grammar use.

Interestingly, no meaningful differences were observed between B2 and C levels on the Aptis speaking test. There are several possible explanations for the lack of significant differences. First, it is possible that there is not a meaningful difference between B2- and C-level performances on the Aptis speaking test; however, such an argument cannot be simply made without exhausting all possibilities that can influence the results of the study.

First, the sample size of this study is not large. Although different statistical procedures were employed to reduce Type I error, the sample size might not be sufficiently large to produce enough statistical power to detect differences between B2 and C-levels.

Second, the array of performance features examined in this study, though comprehensive, is not exhaustive. There might be other features that can characterise the difference between B2 and C level performances. One example of another feature is the precision and register/style of language use. While lexico-grammatical complexity and accuracy can effectively distinguish performance among lower levels on the CEFR, performance at the higher level might be characterised by the use of more precise language to differentiate finer shades of meaning (see, e.g., Council of Europe, 2001; for the descriptors for C1 and C2 levels on the CEFR). For example, under qualitative aspects of spoken language use on the CEFR, the descriptor for B2 level states, "has a sufficient range of language to be able to *give clear descriptions, express viewpoints on most general topics*, without much conspicuous searching for words, using some complex sentence forms to do so". In contrast, at the C1 and C2 levels, the descriptors state "has a good command of a broad range of language allowing him/her to select a formulation to *express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say*", and "shows great *flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely*, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms", respectively (p. 28). These descriptors suggest that at the higher levels, lexico-grammar use is more characterised by the efficiency and precision of language use, rather than the breadth of lexico-grammatical knowledge of the examinees. Thus, the commonly used accuracy and complexity features (e.g., those used in this study) might not be sensitive to capture the nuance of language use.

Third, the lack of meaningful differences between B2 and C levels might also be because the Aptis speaking tasks only target B1 and B2 level, not above. Therefore, the tasks might not be designed to capture the differences between B2 and C levels well.

Table 16: *CAF features characterising the differences across CEFR levels*

	A2 – A1	B1 – A2	B2 – B1	C – B2
Automaticity of lexico-grammar use	+	+	+	-
Macro-level fluency features	-	+	+	-
Global judgments of speech quality	+	+	-	-
Pronunciation-short vowels	-	-	-	-
Pronunciation-long vowels	-	-	-	-

6. CONCLUSIONS AND IMPLICATIONS

This project employed a corpus-based approach to investigate the CAF features of speaking performances on the Aptis test across different CEFR levels, as an effort to examine the criterion-related and cognitive validity evidence for the test. An array of CAF features were examined, which can be classified into six sub-components: lexical sophistication, lexical appropriateness, grammatical complexity, grammatical accuracy, fluency and pronunciation. The results of this project reveal distinguishing features in all three CAF components. However, post-hoc comparisons suggest that Aptis speaking performances at different CEFR levels are characterised by different CAF components. Interestingly, while lower proficiency levels can be distinguished by more CAF features, no meaningful differences in CAF features were observed between higher proficiency levels.

Findings of this study provide supporting evidence for criterion-related and cognitive validity for the Aptis speaking test. In terms of criterion-related validity, the majority of the CAF features show moderate to strong correlations with CEFR levels. These relationships align with findings of previous research of CAF features in language tests. The PCAs yielded four components that largely correspond to the scoring criteria for the Aptis speaking test. These components, covering lexico-grammatical knowledge, automaticity in language use, macro-level speech fluency, and pronunciation, are also included in CEFR descriptors for speaking ability. This indicates the alignment between key criteria assessed in Aptis and components of speaking ability on the CEFR. Except for pronunciation dimensions, all components showed moderately strong to strong correlations with CEFR levels, suggesting that overall, the rating criteria reflect the systematic differences across proficiency levels on the Aptis speaking test.

In this connection, findings of this study can also be used to assist in rater certification, calibration and scoring procedures for the test. For instance, co-occurring CAF features in benchmark speaking performances can be emphasised when training raters to align to the rating scale. Raters can be trained to pay attention to features related to automaticity of lexico-grammar use to distinguish A1 and A2, and features related to macro-level fluency to distinguish A2 and B1 levels.

In terms of cognitive validity, the findings regarding micro-level fluency features add cognitive validity evidence to Aptis speaking test by formulating possible explanations to the occurrence and recovery of disfluency features across CEFR levels. Specifically, higher proficiency test-takers tend to be more capable of using silent pauses as a means to facilitate sentence parsing and formulation of content while maintaining the flow or smoothness of speech. When unexpected pauses occur, higher proficiency learners tend to be more capable of repairing the disfluency by supplying the appropriate lexico-grammatical items to sustain the meaning or topic of the utterance before the pauses; in contrast, lower-proficiency test-takers tend to show failed attempts in lexico-grammatical repair and are more likely to abandon the topic of the original utterance. These failed attempts might create processing difficulty for the listener, thus resulting in lower intelligibility and comprehensibility.
At a theoretical level, the strong correlations for micro-level fluency features suggest that speech fluency should not only be viewed as amount, rate, and pausing features; it also reflects the cognitive processes of speech production, connecting temporal features with the automaticity in lexico-grammar use.

In this connection, the findings of this project stand to have useful implications for research on L2 speaking assessment and speech fluency in general.

REFERENCES

Anderson, Hsieh, J., Johnson, R. & Koehler, K. (1992). The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*(4), 529–555.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Biber, D. (1988). *Variation across speech and writing.* Cambridge, England: Cambridge University Press.

Biber, D., Gray, B. & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5–35.

Boersma, P. & Weenink, D. (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.05, retrieved 24 January 2014 from http://www.praat.org/.

Bosker, H. R., Quené, H., Sanders, T. & de Jong, N. H. (2014). Native 'ums' elicit prediction of low-frequency referents, but non-native 'ums' do not. *Journal of memory and language*, *75*, 104–116.

Clark, H. H. & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84,* 73–111.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press / Council of Europe.

Corley, M. & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*(4), 589–602.

Cunningham, J., Nicol, T., King, C., Zecket, S. & Kraus, N. (2002). Effects of noise and cue enhancement on neural responses to speech in auditory midbrain, thalamus and cortex. *Hearing Research*, 169, 97–111.

de Jong, N. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods, 41*(2), 385–390.

Derwing, T. M., Rossiter, M. J., Munro, M. J. & Thomson, R. I. (2004). Second Language Fluency: Judgments on Different Tasks. *Language Learning, 54*, 655–679.

Dornyei, Z. & Kormos, J. (1998). Problem-solving mechanisms in L2 communication. *Studies in Second Language Acquisition, 20*, 349–385.

Ellis, R. (2008). *The study of second language acquisition.* Oxford University Press.

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

Housen, A. & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461–473.

Isaacs, T. & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34,* 475–505.

Jenkins, J. (2000). *The Phonology of English as an International Language: New Models, New Norms, New Goals*. Oxford: Oxford University Press.

Kang, O., Rubin, D., Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal, 94,* 554–566.

Lenth, R. (2016). Least-Squares Means: The R Package lumens. *Journal of Statistical Software*, 69(1), 1–33.

McNamara, D. S., Graesser, A. C., McCarthy, P. M. & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

McCarthy, P.M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment, *Behavior Research Methods, 42*(2). 381–392.

O'Sullivan, B. & Weir, C. J. (2011). Test development and validation. In O'Sullivan, B. (Ed.) *Language Testing: Theories and Practices* (pp. 13–32), Basingstoke: Palgrave Macmillan.

Rosenfelder, I., Fruehwald, J., Evening, K. & Yuan, Jiahong. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite. http://fave.ling.upenn.edu.

RStudio Team (2016). *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA. URL: http://www.rstudio.com/.

Saito, K., Trofimovich, P. & Isaacs, T. (2015). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics.* Published online 2 February 2015.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Stevens, K. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Templin, M. (1957). *Certain language skills in children*. Minneapolis: University of Minnesota Press.

Trofimovich, P. & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15,* 905–916.

Weir, C. J. (2005). *Language Testing and Validation: an Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

APPENDIX 1:
Questionnaire for global judgment of speech

Please use the slider scale to indicate to what extent you agree with the following statements.
On this slider scale, 1 represents *strongly disagree*, and 5 represents *strongly agree*.

1. The speaker's speech is intelligible.
2. The speaker's speech is comprehensible.
3. The speaker's speech is lexically sophisticated.
4. The use of vocabulary is appropriate.
5. The speaker's speech is grammatically complex.
6. The speaker is fluent.

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

COMPLEXITY, ACCURACY AND FLUENCY (CAF) FEATURES OF SPEAKING PERFORMANCES ON APTIS ACROSS DIFFERENT LEVELS ON THE COMMON EUROPEAN FRAMEWORK OF REFERENCE (CEFR)

AR-G/2018/1

Xun Yan
Ha Ram Kim
Ji Young Kim

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203