Language Testing in Asia

RESEARCH Open Access

Exploring the relationships between ASS indices and CAF and the impact on Chinese college students' oral English performance

Xiaoqin Shi[1*†], Xiaoqing Wang[1†] and Wei Zhang[1]

†Xiaoqin Shi and Xiaoqing Wang contributed equally to this work.

*Correspondence: 1135014997@qq.com

[1] College of Language and Culture, Northwest A&F University, Yangling 712100, Shaanxi Province, China

Abstract

Automatic Speech Scoring (ASS) has increasingly become a useful tool in oral proficiency testing for Second Language (L2) learners. However, limited studies investigate the alignment of ASS indices with the Complexity, Accuracy, and Fluency (CAF)—the three dimensions in evaluating L2 speakers' oral proficiency, and the subsequent impact indices on the oral performance of Chinese college students. To bridge this gap, this study used comparative analysis, Pearson analysis, and linear regression analysis to delve into the relationship and correlations between paired ASS indicators of "pronunciation", "fluency", "integrity", "speed", "duration", and "overall", while also analyzing the relationships between "overall" and other variables. These analyses were conducted using 956 audio clips of freshmen who took the College English Test-Spoken English Test Band 4 (CET-SET-4) in May 2022 in China. The findings reveal that (1) the ASS indicators and evaluation methods are similar but not identical to those employed in prior studies; (2) "pronunciation" encapsulates both the accuracy and fluency dimensions of CAF; (3) "pronunciation" and "integrity" have significant impacts on Chinese college students' oral English performance in read-aloud tasks. The study suggests that future research should further investigate the specific pronunciation challenges faced by Chinese college students, such as phonetics, stress, and intonation. Additionally, it highlights the need to comprehend teachers' attitudes and preferences towards ASS to enhance its effectiveness in assessing second language (L2) learners' oral proficiency. The study would provide some references to teachers for oral English teaching design and students for their self-assessment of oral English proficiencies.

Keywords: ASS, L2 Chinese college students, CAF, Oral performance evaluation

Introduction

In second language (L2) learning, pronunciation serves as a crucial indicator of oral proficiency. L2 Learners aspire to achieve a pronunciation level close to that of native speakers, reflecting their dedication and aspirations in language acquisition (Saito, 2019; Suzukida, 2021). This aspiration has sparked scholars' interest in developing suitable metrics and instruments to measure L2 learners' pronunciation abilities.

Traditionally, L2 learners' oral performance was evaluated by human raters based on standard native pronunciation and phonetic rules (Higgs & Clifford, 1982).

However, CAF (complexity, accuracy, and fluency) has emerged as three key dimensions for assessing L2 learners' oral proficiency and globally gained widespread acceptance among scholars (Housen, 2021; Housen & Kuiken, 2009; Vercellotti, 2015; Yuan & Ellis, 2003). Housen (2021) emphasized that every dimension of CAF has its subcomponents like the stress of pronunciation is a subcomponent of the fluency dimension. Vercellotti (2015) used dynamic system theory to gauge CAF. In pursuit of quantifying, researchers have sought out various tools and software to measure CAF. The speech analysis software, Pratt, has gained popularity for analyzing audio features in speeches, thereby enhancing the objectivity of L2 oral assessments (Lahmann, et al., 2017; Sabu & Rao, 2018; Uddin & Nilsson, 2020). Although they did plenty of research about CAF, few scholars have made it clear which dimension of CAF pronunciation should be attributed to.

With the rapid advancements in AI technology, Automatic Speech Scoring (ASS) system has risen for oral tests. This system boasts multi-model task operations and high objectivity in evaluating oral proficiency (Bamdev et al., 2023; Bhat & Yoon, 2015; Zechner, 2009). Bamdev (2023) underscored that ASS can score speech features, including pronunciation and fluency. Xu, et al (2021) emphasized that ASS can give more lenience and fairness in scoring low-proficiency speakers. Scholars have paid more attention to ASS's algorithms and their scoring reliability, seldom considering ASS indices and their internal correlations.

In China, ASS has been effectively utilized to improve the English pronunciation of college students by identifying pronunciation and stress errors. Liu, et al. (2021) emphasized that ASS can offer learners more detailed feedback on pronunciation errors, while Fouz-González (2020) confirmed that ASS could provide learners with standard pronunciation models. However, the underlying reasons why ASS can evaluate learners' oral performance and the distinction between ASS and the widely used CAF indexes remain underdeveloped. By addressing these gaps, this study will subdivide the ASS assessment indices and delve into the relationship between ASS and CAF, as well as explore the internal relationships among ASS indexes within a specific test task. Our study will enhance students' understanding of ASS functions and provide scientific guidance for L2 oral teaching and assessment.

Literature reviews

CAF measurements of L2 learners' oral performance

To assess the oral performance of L2 learners, Higgs and Clifford (1982) proposed a five-factor framework, which includes vocabulary, grammar, pronunciation, fluency, and accuracy. This framework laid the foundation for a contemporary three-dimensional CAF assessment model. Then, various indices have been customized to operationalize the three dimensions. Concerning complexity, Yuan and Ellis (2003) argued that speech complexity could be assessed by syntactic complexity and diversity. Furthermore, the terminal unit (T-unit), or clause unit (C-unit) has become a commonly used metric for complexity (Malicka, 2020; Mora & Valls-Ferrer, 2012; Norris & Ortega, 2009).

As for accuracy, Iwashita et al. (2008) pointed out that meaningful words and target-like syllables serve as crucial elements for assessing pronunciation accuracy. Kuang (2017) took a more technical approach by measuring intonation accuracy

through the analysis of pitch and fundamental frequency (F0) using Praat software. Tremblay et al. (2018) pointed out that the connection between words, or word liaison, can reveal errors in spoken language. These studies placed a strong focus on pronunciation accuracy.

Turning to fluency, Skehan (2003) and Segalowitz (2010) both emphasized the pivotal role of utterance fluency in determining L2 learners' oral proficiency. Given that the prosody of non-native speakers can significantly affect native-speaking listeners' comprehension, speech flow has been a focal point in previous studies. For example, Lekwilal (2021) measured the rises and falls in pitches, as well as the correct and incorrect pause locations in participants' speeches to gauge speech flow, as well as the ratio of syllable or pause numbers and speech running time (De-Jang, et al. 2015; Prefontaine, 2015).

Additionally, for the measurement of speech speed, researchers calculated factors such as the number of syllables per second (Ellis, 2009; Nagy & Brook, 2020), syllable duration (Lahmann et al., 2017), and the running time of a speech (Zhang et al. 2001). More recently, researchers measured both filled and unfilled pauses, as well as inter- or intra-sentence pauses (Kirjavainen et al., 2022; Kosmala & Cribe, 2023). Based on the studies, it is evident that the complexity dimension of CAF has been overlooked, as the complexity measurements scarcely pertain to pronunciation.

Pronunciation measurement using Auto Speech Scoring

Around two decades ago, the first Automatic Speech Scoring (ASS) system emerged, initially designed to rate the pronunciation of native speakers (Bernstein et al., 1990). Subsequently, its applications broadened to encompass the assessment of non-native speakers' speech (Bernstein, 1999), primarily focusing on read-aloud and repetitive-read tasks (Evanini et al., 2017). Leveraging Automatic Speech Recognition (ASR), ASS detects errors within a predefined text set (Witt & Young, 2000) and assigns scores to factors such as fluency, prosody, intonation, stress, and vocabulary use. ((Evanini et al., 2017; Zechner et al., 2009). Scholars have also verified the alignment between ASS and human rater assessment, including the ASS model's adoption and scoring procedures. Prominent ASS systems include Speech-Rater, developed by Educational Testing Service (EST) in the United States, and Versant, created by Pearson. Both systems are widely used to assess the oral proficiency of non-native speakers (Jiang & Chen, 2021; Sun, 2021). Bernstein and Cheng (2008) employed Versant to evaluate 159 speeches and found a significant correlation between ASS assessment and human rater scores on fluency and accuracy ($r = 0.75–0.94$). However, the test's limited focus on sentence proficiency and vocabulary as proxies for discourse content makes it less effective in addressing advanced measures such as discourse organization and viewpoint organization. In contrast, the SpeechRater system excels in directly assessing candidates' natural speech. Chen et al. (2018) used SpeechRater to evaluate content accuracy (e.g., N-element model, sequence matching, and fixed expression.), speech style (e.g., fluency, prosody, and pronunciation), and language use (e.g., vocabulary and grammar) in TOFEL iBT. This test specifically gauges a test taker's ability to articulate his thoughts verbally in an academic context. Regardless, these systems rely on unambiguous

definitions of measurement objectives and constructs, providing a solid foundation for the development of future spoken language testing tools.

Oral proficiency evaluation of Chinese L2 learners

Over the past few decades, the Oral Proficiency Interview (OPI) has been the primary method used to evaluate the speaking ability of Chinese learners. It involves a panel of interviewers supported by second-raters for scoring. previously, the Pratt test was widely used to measure the pronunciation accuracy of L2 interpreters' speech (Bai, 2022) and to measure syllable stress by pinpointing vowels and consonant positions within words (Xie, 2019). However, earlier studies were inevitably constrained by limited sample size due to substantial workload. Nowadays, the number of test papers scored by ASS is virtually unlimited, which has contributed to its growing popularity among scholars. In the past decade, ASR technologies have gained acceptance as assistants to raters in large-scale, high-stakes tests like CET-SET-4 or- 6 (Gong, et al. 2009), as well as tests involving read-aloud or repetitive tasks (e.g., Li, et al. 2008; Li & Yan, 2012). While the consistency coefficient (r) between machine-rater scoring ranges from 0.6 to 0.9, none of these studies have explored the relationship between the paired indexes of ASS. Additionally, they failed to explore a predictive model that could accurately forecast participants' oral performance when assessed by ASS.

Research questions

The consensus among observers is that Chinese college students tend to have a low-intermediate level of oral proficiency (Jiang and Dai, 2018; Yu, 2020). Students often lack sufficient feedback on their fluency and pronunciation errors, and they struggle to access synchronized, detailed comments from their teachers and raters. The ASS system can compensate for these drawbacks by providing systematic and meticulous scoring on each metric within its framework. Nevertheless, to the best of our knowledge, there is scant research exploring the underlying reasons for ASS's ability to grade participants' speeches and the relationship between ASS indices and CAF. Furthermore, the relationship between paired ASS indices remains elusive. Consequently, the following research questions arise:

1. What are the relationships between ASS and CAF indices for assessing L2 learners' oral performance?
2. What are the relationships between paired indices of ASS?
3. Which indices of ASS can serve as key predictors of Chinese L2 learners' oral performance in a read-aloud task?

Methodology

In our study, we utilized comparative and quantitative analyses to obtain results. We showcased the pivotal role of the ASS in evaluating the spoken language proficiency of test takers by delving into its working principle and algorithm. Furthermore, we compared ASS indices with those previously used by scholars, and employed Pearson

and regression analyses to obtain corresponding results. In addition, we developed Python modules to format our data and establish a connection with the online ASS server for evaluating audio samples.

The working flow of ASS

Despite slight differences in their internal algorithms, ASS systems adhere to comparable scoring methodologies. The component architecture comprises the speech recognition engine, the backstage confirmation system, the evaluation model, the training database, and the evaluation interface (Fig. 1).

The diagram in Fig. 1 visually displays the essential components highlighted in bold that constitute the ASS, along with a detailed depiction of how it operates in a series of steps. Specifically, the workflow of an ASS system can be outlined as follows: (1) users produce speech based on a provided text; (2) The cloud/platform uploads this audio to the speech recognition engine, which decodes and calculates the audio, through the evaluation interface; (3) the backstage configuration system, which divides the given texts into separate words or annotates sounds/phonemes, stores them, providing alignment standards for the speech evaluation engine. (4) The training database (corpus) forms a speech evaluation model for the engine, serving as the benchmark to obtain the evaluation result through decoding and computational processing; (5) the evaluation results return to the user via the evaluation interface.

The algorithms of ASS

The comprehensive scoring algorithm involves a series of intricate steps aimed at evaluating speech. The total scoring algorithm proceeds with the following steps: (1) Speech feature extraction; (2) reference text alignment; (3) speech-recognition rate calculation;

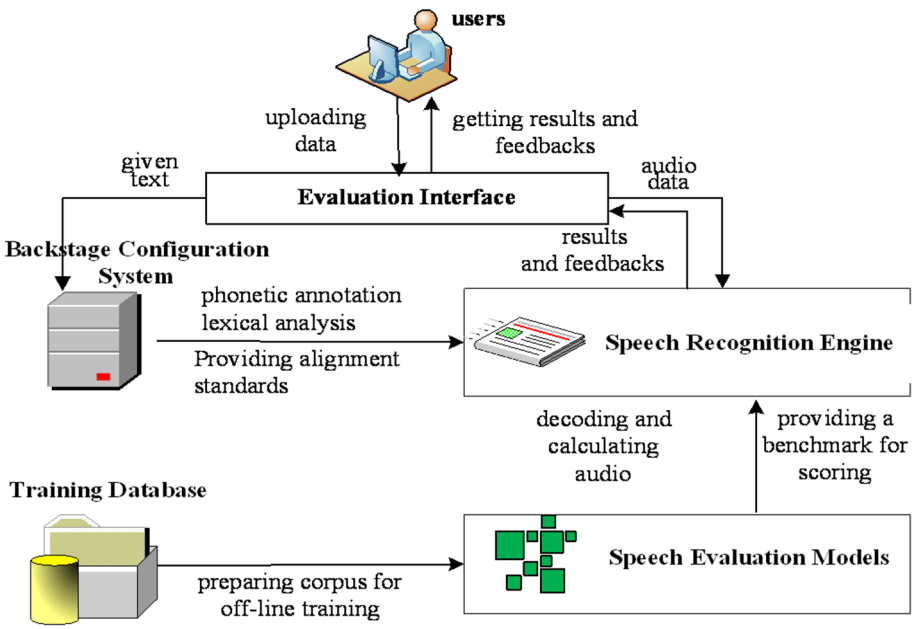

Fig. 1 The architecture of ASS workflow

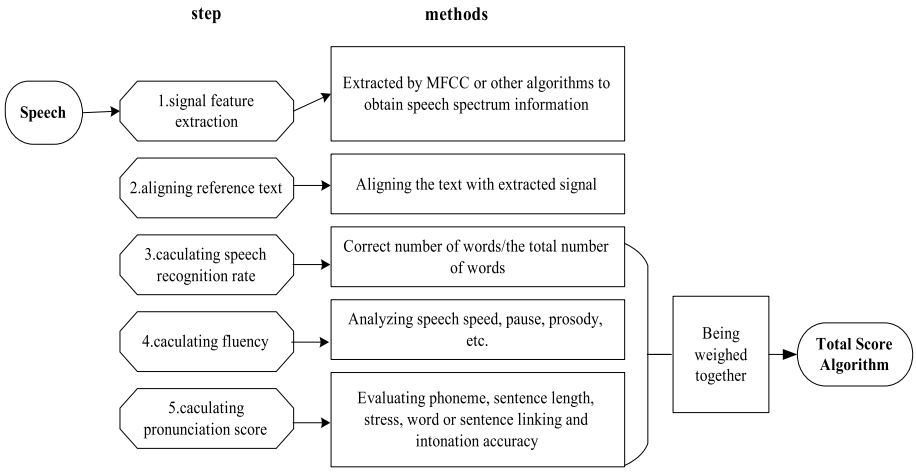

Fig. 2 The frame diagram of the total score algorithm

(4) fluency scoring; (5) Pronunciation scoring; and (6) total score calculation. The specific steps and method descriptions are elucidated in Fig. 2.

Figure 2 delineates the process by which ASS scores the speech step-by-step using specific algorithms, resulting in a total score. This portrayal of steps and methods offers readers a comprehensive understanding of the indices of ASS.

Participants

One thousand fifty-six students from a prestigious university in mainland China were selected to participate in the research study. The students were informed about the study's objectives during an oral English class, where they expressed interest and agreed to use their test audio. All participants were freshmen, 18 years old, and non-English majors. To assess their spoken English skills, they took the College English Test-Spoken English Test-Band 4 (CET-SET-4), a national English exam for college students in mainland China, in early 2022. Before taking the test, they were made aware that their audio recordings would be utilized in a study where their audio would be evaluated, and provided their consent. Additionally, the university's ethical committee issued an ethical confirmation letter.

Instruments

There are three kinds of instruments as follows:

1 Our data collection instrument is Xuefei speech test software, including a server and clients. Its workflow is explained in the "Data collection" section.
2 Our data scoring instrument is Youdao's ASS cloud services.
3 Our auxiliary instruments are several modules developed in Python by the first author.

In China, Xunfei, Tencent, and Youdao are the leading ASS cloud service providers. We opted for Youdao's ASS cloud service as our scoring tool due to its competitive pricing and well-documented API (Application Program Interface). Its workflow and

algorithms are shown in the "The working flow of ASS" and "The algorithms of ASS" sections (see Figs. 1 and 2). To effectively utilize the Youdao ASS cloud service in our study, five modules were developed in Python, including (1) speech format conversion from MP3 to WAV; (2) speech channel and sampling rate conversion; (3) speech cutting into one-minute segments; (4) the interface connecting to Youdao ASS cloud server; and (5) output conversion. Given the complexity of the JSON (JavaScript Object Notation) format of stored data returned by the server, we further converted it to Excel format to enhance comprehension among students. Our modules could be valuable to other educators and evaluators for their teaching and studies. The scoring results retrieved from the cloud server cover indices of pronunciation accuracy at the phoneme, word, and sentence levels, fluency, integrity, speaking speed, overall scores, false or true stress, speech start and end timestamps, and feedback.

Data collection

In China, the Xunfei speech testing software is utilized in the CET-SET-4. The software comprises a server and clients. The server distributes the test questions and collects the audio data from the clients over the network. Test-takers use the clients to record their speeches, and the clients then send the data to the server via the network. The first author operated the server during the test, allowing access to the data copied from the server.

The audio data was sourced from a read-aloud task of CET-SET-4. In the first section of the test, the participants were instructed to read aloud the same sentence, primarily to assess their pronunciation accuracy, fluency, and encompassing factors such as speech flow and speed. We successfully and randomly collected 1056 sound clips from this section and processed them using the five modules we developed in Pycharm (an integrated development environment). The results were then compiled in Excel format. Subsequently, we eliminated the sound clips in which the pronunciation score was "0" or the clips in which the speaking duration was less than 2 s, indicating that the candidate did not speak or spoke fewer than 3 words. After this rigorous filtering process, we retained 965 data pieces for statistical analysis. Based on index integration, six variables (i.e., "pronunciation", "fluency", "speed", "integrity" "duration", and "overall") were categorized into the dimensions of CAF.

ASS indexes explanation

It is important to note that in this study, "pronunciation" is a comprehensive index providing the evaluation of phoneme, stress, word or sentence linking, and intonation accuracy (see Fig. 2). "Fluency" refers to pause and prosody. "Integrity" and "speed" are related to speech recognition rate and speech rate, respectively. The "overall" is an overall score for the audio data of one case. The values of these five variables are retrieved directly from the cloud server. Additionally, the values of "duration" are calculated by subtracting the audio's start time from its end time. A longer duration indicates more frequent pauses during the test.

Data analysis

In our study, we employed comparative analysis to explore the connections between ASS indexes and CAF indices for assessing L2 learners' oral performance (RQ1). CAF

comprises three key dimensions for evaluating EFL learners' oral proficiency. Our primary focus was on the read-aloud task, which has limited relevance to complexity, leading us to exclude the complexity dimension from our considerations. Prior research has demonstrated significant consistency between ASS and human scoring in assessing participants' oral tests (Gong et al., 2009; Zechner et al. 2015). To further investigate this consistency, we sought to compare the indexes and measuring methods of ASS with those of CAF and categorized ASS indexes into CAF dimensions.

Additionally, we conducted statistical analysis in SPSS 27.0. For instance, we explored the correlations between the six variables of ASS indexes ("pronunciation," "fluency," "integrity," "speed," "integrity," "duration," and "overall") using Pearson's analysis to address RQ2 (What are the relationships between paired indices of ASS?).

Finally, for RQ3 (Which indexes of ASS can play key roles in predicting Chinese L2 learners' oral performance in a certain task?), we aimed to find an equation representing the relationship between the dependent variable "overall" and other independent variables. We used linear regression analysis to explore this relationship. During the data analysis, we excluded some variables based on regression analysis prerequisites and found that the three predictors ("pronunciation," "fluency," and "integrity") would affect "overall." We also utilized two model methods ("Stepwise" and "Enter") as the multi-predictors.

Results

The study aims to explore the relationships between indexes of ASS and CAF and the internal correlations between paired indexes of ASS, as well as the main factor which would influence Chinese college students' oral English performance in a read-aloud task. To address our research questions, we conducted data analyses and the results are presented and interpreted as follows.

The relationships between ASS indexes and CAF indices

We revisited the measurements conducted by various scholars (listed in Table 1) on accuracy and fluency dimensions, utilizing tools like Parra to acquire more objective data primarily based on acoustic feature parameters. After a comprehensive analysis, we have not only uncovered similarities but also notable differences between the two methods. The summary of the relationships is presented in Table 1.

As shown in Table 1, the measurements of ASS indexes generally correspond to those cited in prior works, as discussed in the literature review. For instance, Bai (2022) and Xie (2019) gauge syllable stress and pronunciation based on acoustic parameters, like ASS, which measures both by extracting speech features using a certain algorithm. However, certain discrepancies are evident due to the exclusion of ASS to score "word liaison" and "omission of syllables or words" as per the ASS scoring algorithms (depicted in Fig. 2). On the contrary, the ASS scoring parameter "recognition rate", which aligns with the "integrity" index of ASS, is not a factor included in CAF. This congruence and discrepancy will facilitate the improvement of ASS technology and broaden CAF field research.

Table 1 The comparisons between ASS indexes and CAF dimensions

Parameters	ASS variables	CAF indexes	Author(year)	Authors' measurement	CAF dimension
Phoneme	Pronunciation	Syllable pronunciation	Bai (2022)	Extracting the pitch information of both standard English and L2 speech with Praat	Accuracy
Stress	Pronunciation	Syllable stress	Chen (2008) Xie (2019)	In Praat, marking V and C[c)] by spectrum and formant	Accuracy or fluency
Intonation	Pronunciation	Intonation	Kuang (2017)	Tone rise or fall	Accuracy
Word liaison	None	Adjective–noun and noun–adjective	Tremblay et al. (2018)	Eye-tracking movements	Accuracy
Prosody	Fluency	Pitches and formants	Lahmann et al. (2017) Zhang and Wu (2001)	Praat[a)] computing(pause > .25 s) pause > .3 s	Fluency
Pause	Fluency	Inter-sentence pauses (SP)[b)]	Peltonen (2016)	The number of SPs (duration is 0.4 s or longer)	Fluency
Speech flow	Fluency	Speech flow PTR[d)]	De-Jang (2015) Prefontaine (2015)	Syllable number/pause number Phonetic time/speech running time	Fluency
Speed rate	Speed	Syllable number	Nagy (2020) Ellis (2009)	Syllable number per second syllable number/ total run time*60	Fluency
Speech time	Duration	MLR [e)]	Zhang and Wu (2001)	Speech running time	Fluency
Word correction	Integrity	Error-free radio	Jiang and Dai (2018)	The ratio of error-free T-units to total T-units	Accuracy
Omission of syllables	None	A prosodic model	Levey (2002)	Trisyllabic word pairs, Vowel contrast	Accuracy
Recognition rate	Integrity	None			Accuracy
	Overall score: obtained from weighted scores of "pronunciation", "fluency", and "integrity"				

[a] Speech analysis software program

[b] The number of silent pauses (SPs)

[c] *V* Vowel, *C* Consonant

[d] *PTR* Phonetic time ratio

[e] *MLR* Mean length of run

Table 2 The correlations between paired variables

Variables	Statistic	Pronunciation	Fluency	Speed	Integrity	Duration	Overall
Pronunciation	Pearson correlation	1					
Fluency	Pearson correlation	.885**	1				
	Sig	.000					
Speed	Pearson correlation	−0.031	.156**	1			
	Sig	.329	.000				
Integrity	Pearson correlation	.448**	.812**	.340**	1		
	Sig	.000	.000	.000			
Duration	Pearson correlation	.202**	.182**	−.883	.096**	1	
	Sig	.000	.000	.000	.003		
Overall	Pearson correlation	.987**	.949**	0.033	.586**	.200**	1
	Sig	.000	.000	.307	< .001	.000	

** Correlation is significant at the 0.01 level (2-tailed)

The relationships between paired indices of ASS

Each index of ASS is known to contain complex algorithms, but the correlation between these indices has scarcely been investigated. Therefore, this study aimed to explore the relationships between the paired indices via quantitative analysis of our data in SPSS 27.0.and results are shown in Table 2.

As shown in Table 2, "overall" exhibited strong correlations with both "pronunciation" ($r=0.987$, $p<0.01$) and "fluency" ($r=0.949$, $p<0.01$), as evident from their correlation coefficients surpassing 0.7 ($r>0.7$). Meanwhile, it displayed a moderate correlation with "integrity" ($0.3<r<0.7$) and a negligible correlation with "speed" and "duration" ($r<0.3$). Similarly, "pronunciation" strongly correlated with "fluency" ($r=0.885$), moderately with "integrity" ($r=0.448$), and weakly with "duration" ($r=0.202$). On the other hand, "fluency", exhibited relatively weak correlations with both "duration" and "speed" ($r=0.156$ and $r=0.182$, respectively), but a strong correlation with "integrity" ($r=0.812$). Notably, "speed" did not correlate with "pronunciation", but had a weak correlation with "integrity" ($r=0.34$) and a strong correlation with "duration" ($r=-0.883$). Finally, there was virtually no correlation between "integrity" and "duration" ($r=0.096$). It appears that "speed" and "duration" have minimal or no correlation with any other indices. The higher correlations ($r=0.987$, $r=0.949$) suggest that "pronunciation" and "fluency" may be the key indicators of ASS in the read-aloud task of the test.

The key indicators of ASS in the read-aloud task

In "The relationships between paired indices of ASS" section, the findings indicate robust correlations between "overall" and "pronunciation", "overall" and "fluency", and "overall" and "integrity". Furthermore, our objective is to pinpoint the primary influencing factor on "overall" and ascertain if there exists a linear relationship between "overall" and the other three indexes for predicting students' oral proficiency. Our data analysis and results can be found in Table 3.

As can be seen from Table 3, "overall" is highly susceptible to "pronunciation" and "fluency". This conclusion is supported by the Durbin-Watson test, which measures autocorrelation, and its value approximates 2. According to Liu et al. (2003), such an approximation typically indicates a comparatively low level of autocorrelation among the independent variables, suggesting better model adaptability.

Two prerequisites of the regression analysis are (1) residuals follow a normal distribution, and (2) the independent variables are not covariate (Duleba & Olive, 1996). To address these prerequisites, a "Standardized Residual Plots Histogram" was adopted for (1), and "Collinearity Diagnostics" was employed for (2), respectively.

The histogram of residuals (as shown in Fig. 4) validates the Prerequisite (1) of the linear regression. For the Prerequisite (2), the Tolerance (T) values for both "pronunciation" and "fluency" exceeded 0.1, with Variance Inflation Factors (VIF) below 10 ($T=0.217$; $VIF=4.616$). Conversely, "integrity" exhibited a T value lower than 0.1 and a VIF higher than 10 ($T=8.76\text{E}-13$; $VIF=1.14\text{E}+12$). Given the linear correlation between "integrity" and either "pronunciation" or "fluency", it was deemed inappropriate as a predictor for "overall" and it was excluded from the model when the model method is "step-wise". The constant value approximated $1.1*10^{-5}$ (around 0), and the unstandardized coefficients for "pronunciation" and "fluency" were 0.6

Table 3 The linear model and predictors

	Model summary			**Coefficients (dependent variable = "overall")**						
					UN-SC	S C	*t*	Sig	Collinearity diagnostics	
Model method	*R*	R^2	Durbin-Watson	Predictors	*B*	Beta			*T*	VIF
Step-wise	1.0	1.0	1.986	(Constant)	1.10E − 5		6.534	< .001		
				Pronunciation	0.6	0.679	20964425.1	.000	0.217	4.616
				Fluency	0.4	0.347	10715587.8	.000	0.217	4.616
				Integrity (excluded)			3.65	< .001	8.76E − 13	1.14E + 12
	The linear equation: "overall" = 0.6 * "pronunciation" + 0.4 * "fluency" [a]									
Enter	1.0	1.0	2.008	(Constant)	9.89E − 6		5.729	.000		
				Pronunciation	0.80	0.906	5.25E + 7	.000	0.8	1.25
				Integrity	0.20	0.181	1.05E + 7	.000	0.8	1.25
				Fluency (excluded)					2.37E − 13	4.21E + 12
	The linear equation: "overall" = 0.8 * "pronunciation" + 0.2 * "integrity" [b]									

UN-SC Unstandardized coefficients, *SC* Standardized coefficients, *T* Tolerance, *VIF* Variance inflation factor

pronunciation	fluency	integrity	overall	pre_overall_1	pre_overall_2
89.48	94.74	100	91.59	91.58	91.58
80.51	90.26	100	84.41	84.41	84.41
85	88.63	92.25	86.45	86.45	86.45
85.9	92.95	100	88.72	88.72	88.72
89.89	94.95	100	91.92	91.91	91.91
91.57	95.79	100	93.26	93.26	93.26
86.48	93.24	100	89.18	89.18	89.18
79.95	86.1	92.25	82.41	82.41	82.41
90.06	95.03	100	92.05	92.05	92.05
89.81	94.9	100	91.84	91.85	91.85
89.49	94.74	100	91.59	91.59	91.59
65.5	68.48	71.47	66.7	66.69	66.69
90.81	95.41	100	92.65	92.65	92.65
90.23	95.12	100	92.19	92.19	92.18
90.07	95.04	100	92.06	92.06	92.06

Fig. 3 The predicting value vs ASS scoring value of "overall". Note: The values presented in the yellow column are assigned scores by the ASS, while those in the blue column are derived from Eq. 1, and the green column values stem from Eq. 2

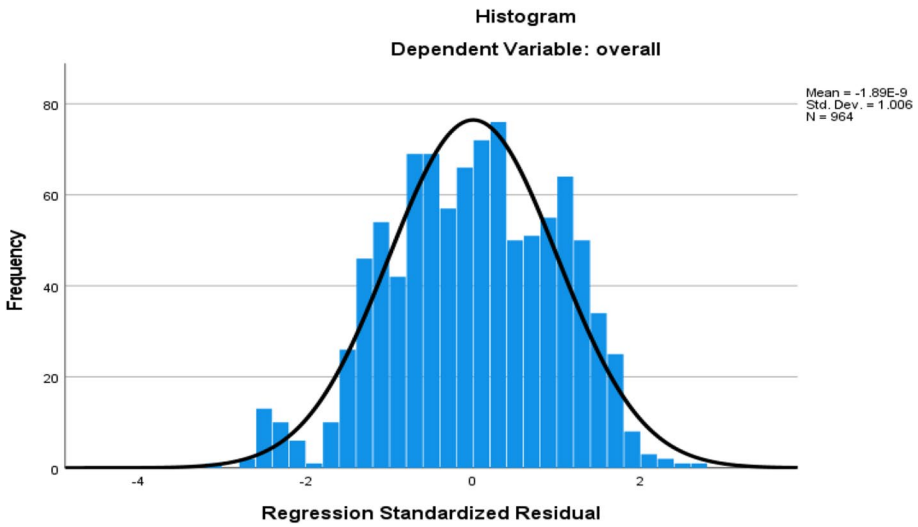

Fig. 4 The histogram of residuals

and 0.4, respectively. The resulting model built among variables was: "overall" = 0.6 * "pronunciation" + 0.4 * "fluency" (designated as Eq. 1). Similarly, under the "Enter" method, the model was built on "overall" = 0.8 * "pronunciation" + 0.2 * "integrity" (called Eq. 2). "Fluency" had a strong correlation with "pronunciation" and "integrity". As a result, "fluency" was excluded since the values of T and VIF indicated a strong cross-correlation of the independent variables ($T = 2.37\text{E} - 13$; $VIF = 4.21\text{E} + 12$).

Figure 3 illustrates the "overall" value alongside the predicted values calculated using Eqs. 1 and 2, providing a visual comparison of the model's predictive capabilities.

Upon inspection of Fig. 3, it appears that there are minimal differences among the three methods used to calculate the "overall" values. However, it becomes evident that Eq. 1 lacks robustness when subjected to more comprehensive data testing.

Based on the data shown in Fig. 4, it is evident that the residuals adhere to a normal distribution. This fulfills a crucial condition for conducting linear regression analysis, indicating that the model is appropriate for predicting the values of the dependent variable.

Discussion

This study first investigated the relationship between dimensions of CAF and indexes of ASS based on various acoustic feature parameters. We discovered that the ASS scores were consistent with the criteria and measurements used by scholars. In contrast to manual scoring conducted by native speaker raters, ASS implements a corpus-based scoring technique that involves extracting acoustic features from loaded speech and subsequently comparing them with standard pronunciations stored in a training database (Bai, 2022). As a result, ASS evaluates pronunciation in a manner like that of scholars. Rogerson-Revell (2021) further echoed this sentiment, highlighting that ASS can compare subjects' pronunciations with those of native speain kers by constructing a model based on a database containing a large number of native speaker speech samples. These pieces of evidence firmly establish the accuracy and objectivity of ASS scoring.

In our study, stress, a key feature of pronunciation in ASS, serves as an index of both accuracy and fluency. This is slightly different from previous studies that solely attribute stress to a single dimension of CAF. Segalowitz (2010) argues that "pronunciation", particularly expressed through stress and intonation, should be categorized under the "fluency" dimension, as perceived fluency refers to the extent to which a speaker's speech can be perceived by the audience. A consensus among Chinese scholars exists regarding the influence of pronunciation stress on the fluency of Chinese students (Liu, 2008; Xia, 2013). Nonetheless, a faction of scholars dissents from this perspective, positing that pronunciation is situated within the "accurate" dimension of the CAF framework (Foster et al., 2000; Iwashita et al., 2008; Liu, 2008). The conflicting classification of pronunciation stress within CAF dimensions underscore the necessity for further research to elucidate its multifaceted role. Additionally, this comprehension can aid in crafting targeted interventions to tackle pronunciation issues, ultimately leading to enhanced communication skills for both native and non-native speakers.

In terms of parameter comparison, the congruity between the ASS indices and the rater evaluation system validates researchers' findings of high consistency ($r > 0.7$) in human–machine speech ratings (Bernstein & Cheng, 2008; Li, et al. 2008; Gong et al., 2009; Zechner et al. 2015). Nevertheless, discrepancies arise when considering the "recognition rate" under the "integrity" index of ASS, which has not been mentioned in previous studies. Furthermore, the "word liaison" and "syllable omissions" as discussed in Tremblay et al. (2018) and Levey (2002) cannot be detected by ASS from Youdao. The differences support Sun's (2021) assertion that SpeechRater (an ASS system) is inadequate for rating participants' oral performance in high-stake tests. This perspective is also consistent with Bridgeman et al.'s (2012) view that SpeechRater cannot measure various aspects of sentence structure (specifically, the complexity dimension of CAF) that can be assessed by human raters. Collectively, the results suggest there are numerous opportunities for future enhancements of ASS, particularly in identifying and processing specific linguistic features such as "word liaison" to improve its effectiveness in language

assessment. Furthermore, this highlights the essential role of human raters in certain assessment scenarios and the potential for a more comprehensive evaluation through a combination of human and automated scoring.

Secondly, our findings underscore the significant correlations between "overall" and three independent indices (i.e., "pronunciation", "fluency" and "integrity") respectively. Also, the correlations emerge among the paired indices, specifically between "pronunciation" and "fluency", "fluency" and "integrity", as well as "speed" and "duration". However, "speed" is observed to have a weak correlation with "fluency", which is inconsistent with previous findings, such as Ellis (2009) and Nagy and Brook (2020). Given that "speed" and "fluency" are independently measured by distinct algorithms within ASS, and "speed" is often considered a variable of "fluency" (Kallio et. al., 2022), our results are not in direct contradiction with previous studies in general. Furthermore, in our study, "speed" and "duration" do not strongly correlate to "overall", which is possible because the read-aloud task chosen for the study was relatively simple, and individual differences were not found to be significant. Interestingly, our study reveals that "pronunciation" cannot be solely attributed to either accuracy or fluency dimension, echoing Skehah and Foster (2012) who argued that an overfocus on accuracy may result in a lack of fluency, and vice versa. It is noteworthy that Skehan and Foster primarily focus on the impact of task conditions on CAF, using clause length to measure accuracy and fluency dimensions. By contrast, the oral task in our study involves a read-aloud task scored by ASS, where fluency scores tend to be much higher due to fewer pauses compared to other task types. These findings further suggest that when test tasks vary, it is imperative to utilize distinct indices for assessment. Additionally, future research should thoroughly investigate which specific pronunciation features significantly influence the improvement of other pronunciation attributes in terms of accuracy or fluency. Thirdly, our study reveals that "pronunciation", "fluency" and "integrity" can be the strongest predictors of Chinese college students' spoken English performance. This aligns with the findings of Higgs and Clifford (1982), who suggested that pronunciation contributed most to low-level students. When assessing the oral proficiency of Chinese students, our study suggests that the "pronunciation" of ASS can correspond to the accuracy of CAF, consistent with those of Jiang's findings (2018), and also in accord with Wang (2015) who advocated that fluency is the most crucial dimension.

According to Table 3, Eq. 1 was discarded due to the linear relationship between "fluency" and "pronunciation", which led to equation instability. Conversely, Eq. 2 remains valid as it predicts "overall" based on "pronunciation" and "integrity". Notably, in our study, "fluency" emerged as the least important factor influencing L2 students' general oral performance. This contrasts with most previous studies that underscored the importance of fluency in assessing EFL learners' oral performance (e.g. Doe, 2021; Segalowitz, 2010). This inconsistency can be explained by several factors. Firstly, our sample derived from a read-aloud task did not account for cognitive fluency. Secondly, there was minimal variation in test-takers' speeches regarding filled or unfilled pauses, a crucial aspect of fluency assessment (García-Amaya, 2023; Kirjavainen, 2022; Kosmala & Cribe, 2023). As a result, the independent sample *t*-test reveals no significant difference in terms of "fluency" in individuals. Consequently,

our findings emphasize that pronunciation and reading integrity significantly influence Chinese students' oral English performance during a read-aloud task. To accurately evaluate participants' oral performance, various indices of CAF should be considered in different test tasks, alongside participants' actual oral proficiency levels. This study's findings have significant implications for educators and assessors, as they can aid in the development of more accurate and comprehensive assessment tools that take into account the various aspects of language proficiency.

In summary, the implications of this study are profound. From a theoretical standpoint, our research offers a novel perspective on L2 language testing by integrating emerging technology into language assessments. This integration not only advances the methodological rigor of testing but also aligns with contemporary theories of language acquisition and assessment. Additionally, by considering pronunciation as a determining factor alongside the established CAF indices, our study sheds new light on the intricate nature of oral proficiency in lower-level L2 learners. From a pedagogical perspective, the findings highlight the importance of considering and utilizing different CAF indices when evaluating L2 learners' oral performance across a range of tasks and proficiency levels. This approach is in line with theories of second language acquisition that recognize the complex interplay of linguistic factors in proficiency development. Furthermore, the study underscores the critical role of teachers in attending to specific pronunciation features, such as tone, intonation, and stress, as these aspects significantly impact the oral proficiency of L2 learners. By incorporating these pronunciation-focused teaching strategies, teachers can foster a more holistic approach to language development that addresses linguistic form and function.

Conclusion

The summary of the study

The study first investigated the relationship between six ASS evaluation indices (i.e., "pronunciation", "fluency", "integrity", "speed", "duration", and "overall") and CAF dimensions for evaluating L2 oral proficiency. Despite minor discrepancies, the ASS indices and measurements of indices well aligned with the previous studies. Notably, the findings of our study suggest that "pronunciation" involves both accuracy and fluency, contradicting the conventional view that classifies it solely as accuracy. Furthermore, our study explored the correlation between paired indices and the potential linear models connecting "overall" with other variables of ASS. The results reveal that the three paired indices were strongly correlated. A stable linear model exists among "overall" (a dependent variable), "pronunciation", and "integrity". In a read-aloud task, it was evident that pronunciation and word integrity exert a significant influence on Chinese college students' oral English performance.

Limitations of the study

While our study aimed to delve into the relevant issues of the ASS assessment indices, it is imperative to acknowledge its limitations to pave the way for future research. Firstly, ASS just scores participants' speeches from a read-aloud task in our study. There is almost no significant difference in individual pauses recognized by ASS. Future studies should aim to collect audio data with longer speeches, such as peer conversations or oral

compositions, to broaden the applicability of the ASS. Secondly, there might be inherent biases in our participant selection, which was mostly composed of freshmen, as well as during the data cleaning process. Despite our efforts to mitigate these biases by random sampling and rigorous deletion of speech clips, future research should continue to refine these processes. Thirdly, although this study took into account variables such as phonemes, stress, and intonation in evaluating "pronunciation" comprehensively, it failed to pinpoint the primary predictor or the root cause of pronunciation errors. Future studies should delve deeper into these aspects. Fourthly, while the study addressed the relationship between paired indices of ASS, it scarcely delved into the relationship between ASS and raters for each index. Further studies are needed to compare ASS and human rater assessments in greater detail and investigate whether ASS can serve as reliable raters for oral English assessment. Lastly, it would be intriguing to understand the attitudes and preferences of teachers and students towards ASS for oral English evaluation. Such insights could inform the further development and application of ASS in educational settings.

Abbreviations

API	Application Program Interface
APP	Application
ASR	Automatic Speech Recognition
ASS	Automatic Speech Scoring
CAF	Complexity, Accuracy, and Fluency
CET-SET-4	College English Test-Spoken English Test Band 4
EST	Educational Testing Service
iBT	Internet-based test
JSON	JavaScript Object Notation
L2	Second language
MP3	MPEG Audio Player 3
OPI	Oral Proficiency Interview
RCM	Relative Contribution Model
TOEFL	The Test of English as a Foreign Language
T	Tolerance
VIF	Variance inflation factors

Acknowledgements
The authors would like to acknowledge co-workers for collecting the audio data. The authors are grateful to some teachers for their help in rating some audio data. The authors also thank students for their contributions to the speech data.

Authors' contributions
Xiaoqin Shi: data collection; data analysis; writing, review, and editing. Xiaoqing Wang: writing, review, and editing. Wei Zhang: data curation. All authors read and approved the final manuscript.

Funding
The study has been supported by the Department of Higher Education of the Ministry of Education of China under Grant (No. 230900960264658).

Availability of data and materials
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 27 February 2024 Accepted: 17 July 2024
Published online: 29 July 2024

References

Bai, J. (2022). The Influence of segment and prosody on the difficulty of phonetic variational English interpretation. *Technology Enhance Foreign Language Education, 03*, 23–28+105. http://cnki.net.

Bamdev, P., Grover, M. S., Singla, Y. K., et al. (2023). Automated Speech Scoring System Under the Lens. *International Journal of Artificial Intelligence in Education, 33*, 119–154. https://doi.org/10.1007/s40593-022-00291-5

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. In ProcICSLP-90: 1990 International Conference on Spoken Language Processing (pp. 1185–1188). Kobe, Japan. https://doi.org/10.21437/ICSLP.1990-313.

Bernstein, J. (1999). *PhonePass testing: structure and construct*. Menlo Park: Ordinate.

Bernstein, J., & Cheng, J. (2008). Logic and validation of a fully automatic spoken English test. In V. M. Holland & F. P. Fisher (Eds.), *The path of speech technologies in computer assisted language learning: From research toward practice*. New York: Routledge.

Bhat, S., & Yoon, S. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication, 67*, 42–57. https://doi.org/10.1016/j.specom.2014.09.005

Bridgeman, B., Trapani, B., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27–40. https://doi.org/10.1080/08957347.2012.635502

Chen, H. (2008). A study on the rhythm classification of stress repetition in English reading. *Foreign Languages and Foreign Language Teaching, 3*, 35–37. https://cnki.net.

Chen, L., Zechner, K., Yoon, S., Evanini, Y. K., et al. (2018). Automated scoring of nonnative speech using the SpeechRater v. 5.0 Engine. *ETS Research Report Series, 1*, 1–31. https://doi.org/10.1002/ets2.12198

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

Doe, T. (2021). Fluency development in an EFL setting: a one-semester study. Language Teaching Research, 1–22. https://doi.org/10.1177/13621688211058520

Duleba, A. J., & Olive, D. L. (1996). Regression analysis and multivariate analysis. *Seminars in Reproductive Endocrinology, 14*(2), 139–153. https://doi.org/10.1055/s-2007-1016322

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

Evanini, K., Maurice, H., & Hakuta, K. (2017). Approaches to automated scoring of speaking for K–12 English language proficiency assessments. *ETS Research Report Series, 1*, 1–11. https://doi.org/10.1002/ets2.12147

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

Fouz-González, J. (2020). Using apps for pronunciation training: an empirical evaluation of the English File Pronunciation App. *Language Learning & Technology, 24*(1), 62–85. 10125/44709.

García-Amaya, L. (2023). Investigating the relation between L2 pauses, syntactic complexity, and pause location: Longitudinal data from L2-Spanish study-abroad learners. Second Language Research, 1–31. https://doi.org/10.1177/02676583231152652

Gong, L., Liang, W., & Ding, Y. (2009). Feasibility analysis and practice research on the adoption of machine marking for large-scale oral English test following reading questions. *Technology Enhanced Foreign Language Education, 2*, 10–15. https://cnki.net.

Higgs, T., & R. Clifford. (1982). The push towards communication. In T. V. Higgs (ed): Curriculum, Competence, and the Foreign Language Teacher. Lincolnwood, IL: National Textbook Company, pp. 57–79. https://api.semanticscholar.org/CorpusID:154599357

Housen, A. (2021). Complexity, accuracy, and fluency (CAF). In H. Mohebbi & C. Coombe (Eds.), *Research Questions in Language Education and Applied Linguistics*. Springer, Cham: Springer Texts in Education. https://doi.org/10.1007/978-3-030-79143-8_136

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

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 29–49. https://doi.org/10.1093/applin/amm017

Jiang, C., & Dai, J. (2018). Correlation analysis between oral accuracy measurement and oral performance of Chinese English learners. *Foreign Language Teaching Theory and Practice, 02*, 37–43. https://cnki.net.

Jiang, J., & Chen, D. (2021). A study of automatic scoring on subjective questions. *Foreign Language in China, 18*(6), 58–64. https://cnki.net.

Kallio, H., Suni, A., & Šimko, J. (2022). Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of English with different language backgrounds. *Language and Speech, 65*(3), 571–597. https://doi.org/10.1177/00238309211040175

Kirjavainen, M., Crible, L., & Beeching, K. (2022). Can filled pauses be represented as linguistic items? *Investigating the Effect of Exposure on the Perception and Production of Um, Language and Speech, 65*(2), 263–289. https://doi.org/10.1177/00238309211011201

Kosmala, L., & Cribe, L. (2023). The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence. *Language and Speech, 65*(1), 216–239. https://doi.org/10.1177/00238309211010862

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America, 142*(3), 1693–1706. https://doi.org/10.1121/1.5003649

Lahmann, C., Steinkrauss, R., & Steinkrauss, M. R. (2017). Speed, breakdown, and repair: An investigation of fluency in long-term second-language speakers of English. *International Journal of Bilingualism, 21*(2), 228–242. https://doi.org/10.1177/1367006915613162

Lekwilal, P. (2021). "Read it like you mean it": developing prosodic reading using reader's theater. *Reflections, 28*(1), 1–18. https://files.eric.ed.gov/fulltext/EJ1296444.pdf.

Levey, S., & Schwartz, R. G. (2002). *Syllable omission by two-year-old children*. . https://doi.org/10.1177/15257401020230040201

Li, M., Yang, X., Fen, G., Wu, M., Chen, J., & Hu, G. (2008). Feasibility study and practice of large-scale college oral english test reading question marking by machine. *Foreign Language World, 4*, 88–95. https://cnki.net.
Li, Y., & Yan, Y. (2012). A study of an automatic scoring system for English-speaking exams with multi-feature fusion. *Journal of Electronics and Information, 9*, 2097–2102. https://cnki.net.
Liu, H. (2021). An empirical study of English learning apps in oral English autonomous learning of higher vocational college English majors. *English Abroad, 16*, 272–273. https://cnki.net.
Liu, Q. (2008). A study on the oral English level of Chinese college students. *Modern Foreign Languages, 01*, 83–89+110. https://cnki.net.
Liu, R. X., Kung, J., Gong, Q., & Hou, X. (2003). Principal component regression analysis with SPSS. *Computer Methods and Programs in Biomedicine, 71*, 141–147. https://doi.org/10.1016/S0169-2607(02)00058-5
Malicka, A. (2020). The role of task sequencing in fluency, accuracy, and complexity: investigating the SSARC model of pedagogic task sequencing. *Language Teaching Research, 24*(5), 642–665. https://doi.org/10.1177/1362168818813668
Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly, 46*, 610–641. https://doi.org/10.1002/tesq.34
Nagy, N., & Brook, M. (2020). Constraints on speech rate: A heritage-language perspective. *International Journal of Bilingualism*. https://doi.org/10.1177/1367006920920935
Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*, 555–578. https://doi.org/10.1093/applin/amp044
Peltonen, P. (2016). Temporal fluency and problem-solving in interaction: An exploratory study of fluency resources in L2 dialogue. *System, 70*, 1–13. https://doi.org/10.1016/j.system.2017.08.009
Préfontaine, Y., & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed method approach. *The Modern Language Journal, 99*(1), 96–112. https://doi.org/10.1111/modl.12186
Rogerson-Revell, P. M. (2021). Computer-assisted pronunciation training (CAPT): Current issues and future directions. *RELC Journal, 52*(1), 189–205. https://doi.org/10.1177/0033688220977406
Sabu, K., & Rao, P. (2018). Detection of prominent words in oral reading by children. *Speech Prosody, 64*, 314–318. https://doi.org/10.21437/SpeechProsody.2018-64
Saito, K. (2019). Corrective feedback and the development of L2 pronunciation. In H. Nassaji & E. Kartchava (Eds.), *The Cambridge handbook of corrective feedback in language learning and teaching*. Cambridge: Cambridge University Press. https://researchgate.net.
Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge. https://doi.org/10.4324/9780203851357
Skehan, P., & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance: a synthesis of the Ealing research. In Housen, A., Kuiken, F., & I. Vedder (eds.) Dimensions of L2 performance and proficiency: complexity, accuracy, and fluency in SLA, 199–220. https://doi.org/10.1075/lllt.32.09fos
Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*, 1–14. https://doi.org/10.1017/S026144480200188X
Sun, H. (2021). A review of automatic marking of spoken English at home and abroad. *Foreign Language Education in China, 4*(2), 28–36. https://cnki.net.
Suzukida, Y. (2021). The contribution of individual differences to L2 pronunciation learning: Insights from research and pedagogical implications. *RELC Journal, 52*(1), 48–61. https://doi.org/10.1177/0033688220987655
Tremblay, A., Spinelli, E., Coughlin, C., & Namjoshi, J. (2018). Syntactic cues take precedence over distributional cues in native and non-native speech segmentation. *Language and Speech, 61*(4), 615–631. https://doi.org/10.1177/0023830918801392
Uddin, Z., & Nilsson, E. G. (2020). Emotion recognition using speech and neural structured learning to facilitate edge intelligence. *Engineering Applications of Artificial Intelligence, 94*, 2–11. https://doi.org/10.1016/j.engappai.2020.103775
Vercellotti, M. L. (2015). The development of complexity, accuracy, and fluency in second language performance: a longitudinal study. *Applied Linguistics, 38*, 90–111. https://doi.org/10.1093/applin/amv002
Wang, H. (2015). The interactive effects of task conditions, oral english output and scores. *Foreign Languages in China, 06*, 65–75. https://cnki.net.
Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication, 30*, 95–108. https://doi.org/10.1016/S0167-6393/2899/2900044-8
Xia, Z. (2013). Prosody realization of topic structure in spoken English learners: an empirical study based on english impromptu speech. *Foreign Language Teaching and Research, 45*(03), 398–410+480-481. https://cnki.net.
Xie, J. (2019). A study on the effect of task complexity on oral output of non-English majors. *Foreign Language Studies, 5*, 64–69. https://cnki.net.
Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: Examining automaker reliability. *Assessment in Education: Principles, Policy & Practice., 28*, 1–26. https://doi.org/10.1080/0969594X.2021.1979467
Yu, H. (2020). Dynamic Development of Chinese learners' oral English fluency – and the interaction among complexity, accuracy and Fluency. *World of Foreign Languages, 02*, 81–89. https://cnki.net.
Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 oral production. *Applied Linguistics, 24*(1), 1–27. https://doi.org/10.1093/applin/24.1.1
Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883–895. https://doi.org/10.1016/j.specom.2009.04.009
Zhang, W., & Wu, X. (2001). A quantitative study on the development of oral fluency as a second language. *Modern Foreign Languages, 04*, 342–351+341. https://cnki.net.
Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., Wang, X. H., & Yoon, S.-Y. (2015). Automated Scoring of Speaking Tasks in the Test of English-for-Teaching (TEFT). *ETS Research Report Series, 2*, 1–17. https://doi.org/10.1002/ets2.12080

Publisher's Note