How Can We Address Rater Subjectivity in English Speaking and Writing Assessment through Analytic Indices?

Hengzhi Hu
Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, Malaysia
https://orcid.org/0000-0001-5232-913X
Harwati Hashim
Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, Malaysia
University Research Group on Eduxcellence: Development of Innovative Curriculum & Pedagogy, Universiti Kebangsaan Malaysia, Bangi, Malaysia
https://orcid.org/0000-0002-8817-427X
Correspondence: harwati@ukm.edu.my
Rokhatoy Abidova
English Language and Literature department, Urgench State University, Urganch, Uzbekistan
https://orcid.org/0000-0002-2035-0177

Abstract

Language assessment often relies on human judgments, particularly for productive skills such as speaking and writing. However, these judgments are prone to variability due to biases, inconsistent application of criteria, and differing perceptions of language quality, raising concerns about fairness and reliability. To enhance objectivity in our institutional English proficiency tests for international students, we implemented analytic complexity, accuracy, and fluency (CAF) indices alongside traditional human ratings as a trial practice. A preliminary implementation, with available data analyzed correlationally, revealed significant variations in the alignment between human ratings and CAF indices, underscoring the need for targeted rater training or reassessment in certain areas. Challenges, such as recruiting skilled coders and establishing interpretive frameworks for CAF scores, were identified from our reflections, underscoring the need for further refinement.

Keywords: language assessment, speaking, writing, English, CAF

1. Introduction: Issues Arising

Language assessment, a critical domain within applied linguistics, plays a pivotal role in evaluating linguistic competence and supporting effective language instruction. However, assessing productive skills such as speaking and writing poses unique challenges due to their reliance on human raters. Unlike receptive skills, which can be assessed through standardized, objective measures, speaking and writing evaluations often hinge on subjective judgments. These judgments are influenced by raters' individual biases (Sundqvist & Sandlund, 2024), inconsistencies in criteria application (Bagheridoust & Khairullah, 2024), and varying perceptions of language quality (Dinçer & Gezegin, 2024), which raise concerns about fairness

and reliability. After all, how can we ensure fairness when ratings might depend on who is assessing the performance?

The issue of subjectivity in language assessment extends beyond high-profile, standardized English proficiency tests (Ginting et al., 2023), such as the Test of English as a Foreign Language or the International English Language Testing System, to include less formal assessments conducted within institutions or classrooms, whether for formative or summative purposes (Suwandi, 2023). While standardized proficiency tests benefit from substantial funding and robust infrastructure, enabling strategies such as detailed rubrics, standardized rater training, and advanced technologies such as automated scoring systems (Hu, Gong, et al., 2025; Sundqvist & Sandlund, 2024), institutional assessments face distinct challenges. Resource constraints and varying levels of training among raters can lead to inconsistent scoring (Shao, 2024), further complicating efforts to ensure equitable and reliable evaluation.

This issue has been particularly pronounced within our institution, where English proficiency assessments for students, particularly international students from non-English-speaking countries, often exhibit variability in scoring, especially in speaking and writing evaluations. Inconsistent interpretations of performance standards and the lack of a standardized rater training framework have further exacerbated concerns about fairness and reliability. These challenges highlight a broader issue in institutional assessments: the difficulty of ensuring objective evaluations within resource-constrained environments. Recognizing these challenges, we have sought to explore innovative methods to enhance objectivity and reliability in our assessment processes. One promising avenue lies in leveraging linguistic measures such as complexity, accuracy, and fluency (CAF) indices (Michel, 2017), which offer quantifiable metrics to evaluate key aspects of language performance.

Defining Complexity, Accuracy, and Fluency in Language Assessment

Language assessment, particularly for productive skills such as speaking and writing, has long grappled with issues of subjectivity. Traditionally, human raters assess these skills, making evaluations vulnerable to individual biases, inconsistent application of criteria, and varied perceptions of language quality (Giraldo, 2023; Kunnan, 2024). In response, researchers have sought more objective, quantifiable metrics to support evaluative processes. Among the most promising of these are the CAF indices.

Complexity refers to the range and sophistication of linguistic structures used in language communication (Bulté & Housen, 2012). This encompasses both syntactic and lexical dimensions. Syntactic complexity involves the organization of words and phrases into well-formed sentences and is often operationalized by examining measures such as the variety and

length of clauses, the use of subordinate structures, and the overall structural diversity in a learner's output (Ginting et al., 2023; Skehan, 2009b). High syntactic complexity indicates the ability to construct multi-layered sentences that convey nuanced meaning, reflecting a deeper grasp of grammatical rules and language organization.

Lexical complexity, on the other hand, focuses on the breadth and depth of vocabulary used in communication (McKee et al., 2000). This includes the diversity of lexical items, the use of less frequent or more precise words, and the overall variation in word choice (Šišková, 2012). Studies have demonstrated that higher levels of complexity are associated with greater communicative competence, as learners who can utilize a wide array of linguistic forms tend to convey more nuanced meanings (Alonso, 2018; Suwandi, 2023; Zhang, 2020).

Accuracy in language assessment focuses on the correctness of the language produced (Michel, 2017). This includes both syntactic accuracy (proper use of grammatical structures) and lexical accuracy (correct word choice and form) (Luoma, 2004). Unlike fluency and complexity, which measure aspects of language production quantity and sophistication, accuracy provides an indication of a learner's control over language rules. In objective assessments, accuracy is typically quantified through error counts or by comparing language output against established grammatical norms (Joo, 2022). Accurate language use is essential for effective communication, and its evaluation can help identify areas where learners may require targeted instruction.

Fluency denotes the ease and speed with which language is produced and is concerned with the flow of speech or writing, including the frequency and duration of pauses, hesitations, and repairs (Michel, 2017). In language assessments, fluency is often evaluated by measuring the rate of speech, the smoothness of delivery, and the ability to maintain a coherent narrative without excessive interruptions (Dinçer & Gezegin, 2024; François & Albakry, 2021). Fluency is indicative of a learner's ability to organize thoughts and express them in real time, reflecting both linguistic proficiency and cognitive processing speed. Although fluency can be influenced by individual speaking styles (Abdullah et al., 2024), it remains a critical factor in ensuring that language production is both effective and engaging.

The integration of CAF indices into language assessment offers a more objective framework for evaluating speaking and writing performances. By quantifying elements of complexity, accuracy, and fluency, educators can better identify specific areas of strength and weakness in learners' language production. For instance, research has shown that high correlational coefficients between CAF measures and human ratings can enhance the reliability

of assessments, thereby mitigating some of the subjectivity inherent in traditional evaluation methods (Michel, 2017).

While the application of CAF indices has significantly advanced language assessment, challenges remain. These include ensuring that objective measures adequately capture the holistic nature of language use (Suwandi, 2023), particularly in contexts where pragmatic or discourse-level competencies are crucial (Green, 2022). Moreover, in resource-constrained institutional settings, implementing robust, standardized measures of CAF may be more challenging than in high-stakes testing environments with substantial infrastructural support (Min et al., 2020). However, given the long-standing debate over the subjectivity of human ratings in language assessments (Ginting et al., 2023; J. Li, 2019; Y. Li, 2025; Xu et al., 2021), CAF indices provide an objective and valid lens for understanding learners' communication proficiency. This inspired us to design and trial a method that combines CAF and human ratings within our context.

2. The Context

The assessment practice discussed in this article originates from a private university in Malaysia, an institution that hosts a diverse student body, including a significant proportion of international students from Asian countries, such as China and Indonesia. As English serves as the primary medium of instruction in most programs, assessing students' English proficiency is integral to ensuring they can effectively engage with academic content and communicate within the university environment.

The university administers English proficiency assessments to incoming international students as part of their placement process. These assessments focus on evaluating key language skills to determine students' readiness for academic challenges and, if necessary, recommend remedial courses to improve their language abilities. However, the variability in scoring, particularly in productive skills such as speaking and writing, has been a persistent issue. This variability often stems from subjective judgments by human raters, despite efforts to provide standardized guidelines and rubrics. Resource constraints and the diverse linguistic backgrounds of both raters and students further complicate the evaluation process. While raters are experienced language instructors, they may not always receive extensive training in applying consistent scoring criteria or addressing their inherent biases. Additionally, students' varying levels of exposure to English and their familiarity with academic discourse contribute to the complexity of assessing proficiency in a fair and reliable manner.

3. The CAF Method

While retaining human ratings for efficiency, we attempted to use CAF indices to quantify students' performances in the aforementioned English-speaking and writing test. The speaking test consisted of three opinion-based monological tasks, providing greater opportunities for discourse analysis compared to other task types, such as dialogues, which involve more interaction (Hu, Mohd Said, et al., 2025; Sundqvist & Sandlund, 2024). As shown in Table 1, we analyzed the transcripts of students' English-speaking test responses using the following methods:

- Firstly, we approached complexity through syntactic complexity, quantified by the mean length of Analysis-of-Speech-units (MLA), mean length of clauses (MLC), and ratio of subordination based on AS-units (RSAS)—representing length, sub-clausal, and subordination levels of syntax, respectively (Norris & Ortega, 2009)—and through lexical complexity, calculated by the Computerized Language Analysis (CLAN) software program as the D-score (McKee et al., 2000). Notably, unlike previous studies or practices that used the T-unit to measure syntactic complexity at the length level (Gordon-Pershey, 2022), we opted for the AS-unit, as it is better suited to spoken discourse and accounts for the fragmented and dynamic nature of speech. Also, instead of using the type-token ratio for lexical complexity (Michel, 2017), we implemented the D-score calculated by conducting multiple trials on groups of randomly selected words (a process reiterated three times), which has been shown to offer a more robust analysis of lexical diversity, particularly in contexts where shorter utterances or fragmented speech are common.
- We prioritized holistic accuracy over isolating specific grammatical features, focusing on the overall correctness and appropriateness of the students' language use. This approach allowed us to evaluate how well students adhered to the grammatical norms of English in their spontaneous speech without being overly restrictive or prescriptive about specific errors (Joo, 2022). With lexical, morphological and syntactic errors considered, following popular English styles, such as American and British English, we quantified speaking accuracy using the percentage of errorfree AS-units (PEAS) and the percentage of errorfree clauses (PEC) (Ellis & Barkhuizen, 2005).
- Different from common practices of using speech rate as an indicator of speech fluency, we adopted two alternative measures: the number of pauses (NP) and the number of repairs (NR) (Skehan, 2009a). Speech rate, while commonly used, can be influenced by individual speaking styles and physiological factors, making it a less accurate measure for analysis (Gordon-Pershey, 2022). In contrast, NP and NR could provide a

more nuanced understanding of fluency by capturing both the hesitation phenomena and self-corrections often observed in spontaneous speech.

For the writing test, we adopted the following analysis methods:

- The T-unit was used instead as the primary unit of analysis for writing syntactic complexity because, compared with the AS-unit, it is better suited for written discourse (Gordon-Pershey, 2022), where sentences are typically more structured and complete. We used the mean length of T-units (MLT), MLC, and ration of subordination based on T-units (RST) to comprehensively evaluate syntactic complexity (Bazerman, 2009; Norris & Ortega, 2009). For lexical complexity, we used the D-score, as it accounts for students' varying levels of English proficiency, which can result in shorter or fragmented writing.
- Likewise, with global accuracy considered, we used the proportion of errorfree T-units (PET) and the PEC as the indices (Bazerman, 2009; Joo, 2022). While PET provides a broader view of grammatical accuracy at the sentence level, PEC allows for a more granular analysis by focusing on the accuracy within individual clauses. This dual approach enables a more comprehensive understanding of students' writing accuracy.
- For writing fluency, a popular index is to calculate the words written by a candidate per minute. However, this can be problematic because it often reflects the individual's typing speed or writing pace rather than their actual linguistic proficiency or ability to construct coherent ideas (Kendall, 2013). Therefore, we measured the average number of words per T-unit (ANWT) (Bazerman, 2009), which emphasizes the quality and coherence of written discourse rather than the sheer volume produced in a limited timeframe.

Table 1

Speaking and writing CAF indices

Skill	Dimension	Index
Speaking	Syntactic Complexity	MLAS (divide the number of tokens, calculated by CLAN, by the number of AS-units)
		MLC (divide the number of tokens, calculated by CLAN, by the number of clauses)
		RSAS (divide the number of clauses by the number of AS-units)
	Lexical Complexity	D-score (calculated by calculated by CLAN)
	Accuracy (lexical, morphological, and syntactic	PEAS (divide the number of errorfree AS-units by the total number of AS-units)
		PEC (divide the number of errorfree clauses by the total number of

	errors considered)	clauses)
	Fluency	NP (divide the number of filled pauses with fillers and unfilled pauses over 250 milliseconds by the speaking time in seconds)
		NR (divide the number of repairs by the speaking time in seconds)
Writing	Syntactic Complexity	MLT (divide the number of tokens, calculated by CLAN, by the number of T-units)
		MLC (divide the number of tokens, calculated by CLAN, by the number of clauses)
		RST (divide the number of clauses by the number of T-units)
	Lexical Complexity	D-score (calculated by the Computerized Language Analysis program)
	Accuracy (lexical, morphological and syntactic errors considered	PET (divide the number of errorfree T-units by the number of T-units)
		PEC (divide the number of errorfree clauses by the total number of clauses)
	Fluency	ANWT (divide the number of words by the number of T-units)

4. How We Used the Method

The quantification of CAF requires rigorous manual coding of the transcripts of students' responses in the speaking and writing tests, especially in the identification of linguistic components such as AS-units, T-units, clauses, errors, pauses, and repairs. Therefore, we implemented a detailed protocol to ensure consistency and accuracy during the coding process, which included coding symbols such as * for an error, *(.)* for an unfilled pause, and *[ ]* for repairs. To maintain reliability, a team of trained coders was engaged in analyzing the transcripts. Each coder underwent a series of training sessions to familiarize themselves with the linguistic components and their identification criteria.

Pilot testing was conducted, during which coders analyzed the same set of transcripts independently, and their results were compared to achieve a high inter-rater reliability. Discrepancies were resolved through discussions, and adjustments were made to the coding guidelines to ensure clarity and consistency. We also employed software tools to assist with specific aspects of the analysis, such as calculating the D-score and the number of tokens using CLAN. However, most aspects of the analysis, particularly those related to syntactic complexity and fluency, required manual input to ensure nuanced and accurate assessments. When coded transcripts were ready, the test administration team further calculated the score of each index by counting the occurrences of each code, represented by unique symbols, and inputting the raw data into a preset scoring template. This template, developed in alignment

with the CAF framework, automated the computation of indices by applying predefined formulas.

These data were primarily used as supplementary information alongside the scores assigned by human examiners based on predetermined marking rubrics (refer to Appendix A), which, in addition to including criteria for CAF, also evaluated pronunciation for the speaking test, coherence for the writing test, and task completion for both. The CAF data were made available to test developers and raters for research, training, and assessment evaluation purposes. For instance, we analyzed recordings and writing samples from all 102 students who participated in one of our recent speaking and writing tests. The tests were marked by separate teams, each consisting of three professional raters, while the CAF indices were calculated concurrently by other teams. Correlational analysis was conducted using R[1], with preliminary results summarized in Table 2.

Table 2

Correlational statistics of CAF indices with human ratings

Test		**syntactic complexity score (given by Speaking Examiner 1)**	**syntactic complexity score (given by Speaking Examiner 2)**	**syntactic complexity score (given by Speaking Examiner 3)**
Speaking	MLAS	.728**	.213	.572*
	MLC	.817**	.495*	.608*
	RSAS	.896**	.211	.342
		lexical complexity score (given by Speaking Examiner 1)	**lexical complexity score (given by Speaking Examiner 2)**	**lexical complexity score (given by Speaking Examiner 3)**
	D-score	.874**	.412	.628*
		accuracy score (given by Speaking Examiner 1)	**accuracy score (given by Speaking Examiner 2)**	**accuracy score (given by Speaking Examiner 3)**
	PEAS	.912**	.899**	.930**
	PEC	.901**	.911**	.893**
		fluency score (given by Speaking Examiner 1)	**fluency score (given by Speaking Examiner 2)**	**fluency score (given by Speaking Examiner 3)**
	NP	-.943**	-.887**	-.942**
	NR	-.918**	-.903**	-.923**
Writing		**syntactic complexity score (given by Writing Examiner 1)**	**syntactic complexity score (given by Writing Examiner 2)**	**syntactic complexity score (given by Writing Examiner 3)**
	MLT	.501*	.870**	.845**

MLC	.431*	.902**	.864**
RST	.563*	.883**	.871**
	lexical complexity score (given by Writing Examiner 1)	**lexical complexity score (given by Writing Examiner 2)**	**lexical complexity score (given by Writing Examiner 3)**
D-score	.872**	.331	.428
	accuracy score (given by Writing Examiner 1)	**accuracy score (given by Writing Examiner 2)**	**accuracy score (given by Writing Examiner 3)**
PET	.903**	.945**	.911**
PEC	.917**	.932**	.921**

* $p < .05$, ** $p < .01$

The data revealed that Speaking Examiner 1's ratings were deemed accurate and objective, as indicated by the strong, significant correlations between their scores and the analytic CAF index scores, with a correlation coefficient above .70 and a significance value below .05 (Bennett et al., 2022). In contrast, Speaking Examiner 2's ratings showed significant correlations for all accuracy and fluency indices ($r > .70$, $p < .01$) but raised concerns for syntactic complexity (particularly MLAS, $r = .21$ and RSAS, $r = .21$) and lexical complexity ($r = .41$) due to non-significant relationships ($p > .05$) with the analytic complexity scores, indicating the need for interventions such as additional training or possible reassignment. Speaking Examiner 3's ratings were generally acceptable, with significant correlations in most indices, except for RSAS ($r = .34$, $p > .05$). However, weaker correlational strength, particularly in complexity indices ($r < .70$) compared with Examiner 1 suggested that further calibration might be required to enhance the consistency and reliability of their evaluations.

For the writing test, an overall observation was that all three examiners provided accurate scores for students' syntactic complexity and accuracy ($p < .05$), particularly in accuracy, which demonstrated very strong correlational coefficients across all examiners ($r > .70$). However, compared with Writing Examiners 2 and 3, who achieved high correlations between analytic syntactic complexity indices and their assigned scores ($r > .70$), Writing Examiner 1 showed only moderate correlations between their marking of syntactic complexity and the analytic syntactic complexity indices ($r < .70$). This suggested that discrepancies in the evaluation of syntactic complexity existed, particularly for Writing Examiner 1, suggesting a need for further rater calibration or training to ensure greater consistency in assessing syntactic complexity across different examiners. Another major concern emerged regarding the marking of lexical complexity, where no significant correlations were found with the ratings provided

by Writing Examiner 2 (r = .33, p > .05) and Writing Examiner 3 (r = .43, p > .05). This discrepancy suggested that the evaluation of lexical complexity might have been inconsistently applied or misunderstood by some examiners, necessitating further interventions such as clarifying the marking rubrics.

One of the most plausible reasons for the inconsistencies, particularly in assessing syntactic and lexical complexity, is the varying levels of training and experience among examiners (Burak, 2018; Tsagari, 2020). Examiners with extensive experience or rigorous training in language assessment may demonstrate greater consistency in aligning their judgments with analytic measures, as was seen with Speaking Examiner 1, whose ratings strongly correlated with all CAF indices. In contrast, Speaking Examiner 2's weak correlations in syntactic and lexical complexity suggest that they may have been less familiar with complexity assessment or lacked specific training in distinguishing between different levels of structural and lexical sophistication. Similarly, Writing Examiner 1's moderate correlation in syntactic complexity could indicate uncertainty in evaluating complex structures, leading to a reliance on more holistic impressions rather than systematic analytic scoring.

Examiners may also be influenced by cognitive biases that affect their judgment when evaluating language performance (Giraldo, 2023). One such bias is the halo effect, where an examiner's perception of a test-taker's overall proficiency influences their rating of individual linguistic components (Noor et al., 2023). For example, if an examiner perceives a test-taker as highly proficient in fluency, they may overestimate their syntactic complexity or lexical richness, even when objective measures do not support such an evaluation. This may explain why Speaking Examiner 2's ratings strongly correlated with fluency and accuracy but failed to align with complexity measures. Similarly, Writing Examiner 2's and Writing Examiner 3's low correlation with lexical complexity indices may reflect a tendency to focus more on grammatical correctness (accuracy) than the sophistication of vocabulary use, leading to inconsistency in applying lexical complexity criteria.

Another critical factor influencing rating discrepancies is how examiners interpret and apply the scoring rubrics. While assessment rubrics are intended to provide clear evaluation criteria, research has shown that raters often apply rubrics inconsistently (Alaamer, 2021; Trinh, 2020), particularly when constructs such as complexity do not have absolute or easily quantifiable benchmarks (Bagheridoust & Khairullah, 2024). Complexity, especially lexical and syntactic complexity, is inherently more difficult to assess than accuracy or fluency because it involves qualitative judgments about structure variety, lexical sophistication, and appropriateness in context rather than simply counting errors (Dinçer & Gezegin, 2024). If

examiners lack explicit guidelines or exemplars that illustrate different levels of complexity, their assessments may be more prone to individual interpretation, leading to inconsistencies. This is evident in the low correlation of Writing Examiner 1's syntactic complexity ratings and Writing Examiners 2 and 3's lexical complexity scores with analytic measures, suggesting that examiners may not have had a clear or uniform understanding of how to evaluate these dimensions.

Syntactic and lexical complexity are among the most challenging linguistic features to assess reliably because they involve multiple overlapping factors. For example, a longer sentence is not necessarily more complex if it relies on coordination rather than subordination, and a higher lexical density does not always equate to lexical sophistication (Bulté & Housen, 2012; Michel, 2017). This ambiguity in defining complexity may lead to inconsistencies among examiners, particularly those who rely more on impressionistic or holistic judgments rather than objective linguistic indicators (Green, 2022; Kunnan, 2024). In the case of Speaking Examiner 3, whose complexity ratings had weaker correlations with analytic measures, this suggests a potential tendency to focus on fluency and coherence rather than on the actual structural diversity of the test-taker's output. Similarly, Writing Examiner 1's moderate correlation in syntactic complexity suggests that they may have focused more on grammatical correctness rather than assessing the depth of structural variation in writing samples.

The findings suggest that examiners may require further calibration and training to enhance the reliability of their assessments, particularly in evaluating complexity. Standardized training procedures, including benchmarking sessions, norm-referenced training, and periodic recalibration exercises (Alshakhi, 2024), could help ensure that raters develop a more consistent approach to assessing syntactic and lexical complexity. Moreover, providing raters with examples of test-taker responses at varying levels of complexity and aligning them with CAF-based analytic scores could improve their ability to distinguish between different levels of structural and lexical sophistication. Test organizers should consider embedding CAF-informed materials into their rater development programs to reduce interpretive ambiguity and foster shared standards. CAF data can also serve as a valuable tool for post-hoc moderation, rubric revision, and task refinement, contributing to long-term quality assurance. Although resource investment is needed, particularly in coder training, technical infrastructure, and interpretive frameworks, the benefits of increased scoring consistency, diagnostic insight, and institutional credibility justify the effort. Ultimately, combining objective linguistic indices with calibrated human judgment offers a more reliable and pedagogically meaningful approach to evaluating learners' speaking and writing performance.

5. Reflection and Conclusion

When implementing this assessment method, we encountered several challenges. Recruiting professional coders capable of accurately analyzing transcripts for analytic CAF analysis proved particularly difficult. The process demands a deep understanding of linguistic structures and meticulous attention to detail, making it challenging to find individuals with the requisite expertise. Coding linguistic components such as AS-units, T-units, clauses, errors, pauses, and repairs is not only resource-intensive but also prone to variability, which can impact the reliability of the data.

Moreover, we attempted to report the analytic CAF scores measured by the indices to our test-takers. When reporting these scores, if multiple indices were used for a single speaking dimension, such as syntactic complexity or accuracy in speaking and writing, we calculated the mean of the relevant indices to provide an overall score. For dimensions with only one index, such as writing fluency, the individual score was reported directly. However, test-takers indicated that they still prioritized human ratings, which were easier to understand based on the provided marking rubrics. Additionally, we have yet to establish a standardized framework for interpreting CAF scores, such as setting cutoff values to categorize students' proficiency levels as low, medium, or high. This lack of clear interpretive guidelines further limits the immediate utility of CAF scores for students, underscoring the need for future efforts to enhance the clarity and practical application of these indices in assessment contexts.

Therefore, while the integration of CAF indices into our assessment practice presents a promising avenue for addressing rater subjectivity, its practical implementation remains fraught with challenges that demand careful consideration. The complexity of recruiting and training skilled coders, coupled with the resource-intensive nature of the process, highlights the need for investment in specialized training programs or technological solutions to streamline coding and reduce variability. Additionally, the disparity in test-takers' understanding of analytic CAF scores compared to human ratings underscores the importance of developing user-friendly frameworks for interpreting these metrics. Establishing clear guidelines, such as proficiency benchmarks or visual aids for score representation, could significantly enhance the accessibility and relevance of CAF data for both students and educators. Ultimately, while this method offers substantial potential for improving fairness and reliability in language assessment, its effectiveness hinges on overcoming these operational and communicative barriers, necessitating ongoing refinement and innovation in its application.

Note

Writing fluency was excluded from the correlational analysis due to challenges in defining it within the marking rubrics and the difficulty examiners faced in assigning a score for writing fluency based solely on students' written responses (Green, 2022).

Conflicts of Interest

The authors declared no conflicts of interest.

Ethical Considerations

Written informed consent was obtained from all participants of the study, and the research was conducted in line with the ethical guidelines set by the institutional research ethics committee.

Funding

This work was funded by the Faculty of Education, Universiti Kebangsaan Malaysia (grant code GG-2024-030).

Acknowledgements

We would like to extend our special thanks to the study participants: the test-takers who willingly provided consent for us to analyze their test responses, as well as the examiners and test teams who dedicated their time to preparing the data for this preliminary study.

References

Abdullah, A. T. H., Netra, I. M., & Hassan, I. (2024). Difficulties faced by undergraduate students in English public speaking at a Malaysian university. *Arab World English Journal*, *15*(1), 269-282. https://doi.org/10.24093/awej/vol15no1.17

Alaamer, R. A. (2021). A theoretical review on the need to use standardized oral assessment rubrics for ESL learners in Saudi Arabia. *English Language Teaching*, *14*(11), 144-150. https://doi.org/10.5539/elt.v14n11p144

Alonso, R. A. (Ed.). (2018). *Speaking in a second language*. John Benjamins.

Alshakhi, A. (2024). Speaking skill assessment instrument validity: An investigation into instructors' perceptions. *Journal of Ecohumanism, 3*(7), 4203-4217. https://doi.org/10.62754/joe.v3i7.4543

Bagheridoust, E., & Khairullah, Y. K. (2024). A comparative-correlative study of test rubrics used as benchmarks in assessing IELTS and TOEFL speaking skills. *English Language Teaching*, *16*(6), 1-14. https://doi.org/10.5539/elt.v16n6p1

Bazerman, C. (Ed.). (2009). *Handbook of research on writing: History, society, school, individual, text*. Routledge.

Bennett, K., Heritage, B., & Allen, P. (2022). *SPSS statistics: A practical guide* (5th ed.). Cengage Learning Australia.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA* (pp. 23-46). John Benjamins.

Burak, M. (2018). Speaking assessment: Impact on training sessions. *World Science*, *12*(2), 44-48. https://doi.org/10.31435/rsglobal_ws/30122018/6275

Dinçer, M. N., & Gezegin, B. B. (2024). Teachers' perspectives on assessing English speaking skills: A post-new exam model investigation. *Journal of Language Research*, *8*(2), 30-43. https://doi.org/10.51726/jlr.1566625

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford University Press.
François, J., & Albakry, M. (2021). Effect of formulaic sequences on fluency of English learners in standardized speaking tests. *Language Learning & Technology*, *25*(2), 26-41. http://hdl.handle.net/10125/73429
Ginting, R. S., Dalimunte, A. A., Dalimunte, M., Kurniati, E. Y., & Adelita, D. (2023). A critical review of IELTS speaking test. *Journal of Linguistics, Literature and Language Teaching*, *9*(2), 138-155. https://doi.org/10.32505/jl3t.v9i2.7161
Giraldo, F. (2023). *Fostering pre-service teachers' language assessment literacy*. Universidad de Caldas.
Gordon-Pershey, M. (2022). *Grammar and syntax: Developing school-age children's oral and written language skills*. Plural Publishing.
Green, A. (2022). *L2 writing assessment: An evolutionary perspective*. Springer International Publishing.
Hu, H., Gong, Q., & Said, N. E. M. (2025). Exploring a decade of research: A systematic review of computer-based English speaking tests. *Forum for Linguistic Studies*, *7*(4), 788-803. https://doi.org/10.30564/fls.v7i4.8978
Hu, H., Said, N. E. M., & Hashim, H. (2025). Human ratings and complexity, accuracy, and fluency (CAF) indices: A Correlational study of a standardised monologic English-speaking test in China. *Sage Open*, *15*(2), 1-13. https://doi.org/10.1177/21582440251343944
Joo, M. (2022). Effects of pre-task and on-line planning on complexity, fluency, and accuracy in computer-based English speaking and writing tests. *Korean Journal of English Language and Linguistics*, *22*, 938-956. https://doi.org/10.15738/kjell.22..202210.938
Kendall, T. (2013). *Speech rate, pause and sociolinguistic variation: Studies in corpus sociophonetics*. Palgrave Macmillan.
Kunnan, A. J. (Ed.). (2024). *The concise companion to language assessment*. Wiley.
Li, J. (2019). An evaluation of IELTS speaking test. *Open Access Library Journal*, *16*(12), Article 97399. https://doi.org/10.4236/oalib.1105935
Li, Y. (2025). Analysis of teaching difficulties and strategies in IELTS speaking. *Overseas English Testing: Pedagogy and Research, 7*(1), 1-7. https://doi.org/10.12677/oetpr.2025.71001
Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.
McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Applied Linguistics, 15*(3), 323-338. https://doi.org/10.1093/llc/15.3.323
Michel, M. (2017). Complexity, accuracy and fluency (CAF). In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 2-38). Routledge.
Min, Y., Li, C., & Wang, X. (2020). Computer based English speaking test based on artifical neural network. *Computer Science & IT Research Journal*, *1*(1), 29-36. https://doi.org/10.51594/csitrj.v1i1.132
Noor, N., Beram, S., Yuet, F. K. C., Gengatharan, K., & Rasidi, M. S. M. (2023). Bias, halo effect and horn effect: A systematic literature review. *International Journal of Academic Research in Business & Social Sciences*, *13*(3), 1116-1140. https://doi.org/10.6007/IJARBSS/v13-i3/16733
Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578. https://doi.org/10.1093/applin/amp044
Shao, L. (2024). Report on English test design. *Journal of Education and Educational Research*, *11*(3), 6-12. https://doi.org/10.54097/f9c63q24

Šišková, Z. (2012). Lexical richness in EFL students' narratives. *Language Studies Working Papers*, *4*, 26-36. https://www.reading.ac.uk/elal/-/media/project/uor-main/schools-departments/elal/lswp/lswp-4/elal_lswp_vol_4_siskova.pdf?la=en&

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

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

Sundqvist, P., & Sandlund, E. (2024). *Testing talk: Ways to assess second language oral proficiency*. Bloomsbury Publishing.

Suwandi, I. N. (2023). The assessment of students' speaking skills in Indonesian classrooms. *International Journal of Social Science*, *2*(6), 3465-3470. https://doi.org/10.53625/ijss.v2i6.6103

Trinh, L. H. (2020). The effectiveness of using scoring rubrics in academic writing to English-majored students. *Vietnam Journal of Education*, *4*(4), 76-82. https://doi.org/10.52296/vje.2020.83

Tsagari, D. (Ed.). (2020). *Language assessment literacy: From theory to practice*. Cambridge Scholars Publishing.

Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in Education: Principles, Policy & Practice 28*(4), 411-436. https://doi.org/10.1080/0969594X.2021.1979467

Zhang, X. (2020). Improve English speaking skills in the "human-computer dialogue" examination. *Asia Pacific Education*, (22), 191-192. https://doi.org/10.12240/j.issn.2095-9214.2020.22.091

Appendix A: Excerpts of Marking Rubrics

Test	Criteria	5 - Excellent	4 - Good	3 - Satisfactory	2 - Weak	1 - Poor
Speaking	**Syntactic Complexity**	Uses a wide range of complex sentence structures accurately and appropriately.	Uses a variety of complex structures with some minor errors.	Uses some complex structures but relies mostly on simple sentences.	Uses mostly simple sentences with limited complexity.	Uses only simple structures with frequent errors that impede comprehension.
	Lexical Complexity	Demonstrates a rich and varied vocabulary, with precise and appropriate word choices.	Uses a varied vocabulary with some precise word choices.	Uses common vocabulary with occasional variation but limited precision.	Uses limited vocabulary with noticeable repetition and occasional misuses.	Uses very basic vocabulary with frequent repetition and inappropriate word choices.

	Accuracy	Highly accurate use of grammatical structures and word forms, with only minor lapses.	Generally accurate, with occasional minor errors in syntax and word choice.	Errors occur in sentence structure and word form but do not significantly hinder communication.	Frequent grammatical errors and word form mistakes that occasionally obscure meaning.	Numerous grammatical errors and word form mistakes that make comprehension difficult.
	Fluency	Speaks effortlessly with a natural rhythm, minimal hesitation, and smooth delivery.	Speaks smoothly with some natural pauses but maintains good flow.	Shows moderate hesitation but is able to continue speaking.	Hesitation and pauses frequently disrupt speech flow.	Speech is very slow, disjointed, and difficult to follow.
Writing	**Syntactic Complexity**	Demonstrates a wide range of complex sentence structures effectively and appropriately.	Uses a variety of complex structures with some minor errors.	Uses some complex structures but relies mostly on simple sentences.	Uses mostly simple sentences with limited complexity.	Uses only simple structures with frequent errors.
	Lexical Complexity	Exhibits a rich and precise vocabulary with appropriate word choices and variety.	Uses a varied vocabulary with mostly appropriate word choices.	Uses common vocabulary with occasional variation but limited precision.	Uses limited vocabulary with noticeable repetition and occasional misuses.	Uses very basic vocabulary with frequent repetition and inappropriate word choices.
	Accuracy	Highly accurate use of grammatical structures and word forms, with only minor lapses.	Generally accurate, with occasional minor errors in syntax and word choice.	Errors occur in sentence structure and word form but do not significantly hinder comprehension.	Frequent grammatical errors and word form mistakes that occasionally obscure meaning.	Numerous grammatical errors and word form mistakes that make comprehension difficult.