EJOSS
EDUCATUM
Journal of Social Sciences
UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY
ISSN 2289-9391 • eISSN 2462-2443

Mapping the Landscape of Artificial Intelligence (AI)-Powered Assessment: A Bibliometric Analysis of Scopus and Web of Science (WoS)

Nurul Ashikin Izhar[1], Yahya Al-Dheleai[2*]

[1] School of Education Studies, Universiti Sains Malaysia, 11700 Gelugor, Pulau Pinang, Malaysia
[2] Digital Buddy International B.V. Parallelweg 30, Den Bosch, The Netherlands

***Corresponding author email: yamohd3@gmail.com**

ARTICLE HISTORY
Received: 18 August 2025
Revised: 28 November 2025
Accepted: 15 December 2025
Publisher: 20 January 2026

KEYWORDS
artificial intelligence
AI-powered assessment
assessment
bibliometric analysis
BLR

ABSTRACT - The surge of interest in Artificial Intelligence (AI) in higher education has led to rapid growth in research regarding its potential applications in assessment. This study analyzes publication trends, document types, and citation patterns related to AI in educational assessment, alongside the emergence of AI-powered literature search platforms. Data from Scopus and Web of Science (WoS) databases (2020–2024) were retrieved for analysis. A total of 42 publications were analyzed using VOSviewer for keyword mapping and cluster identification, while Harzing's Publish or Perish was utilized for citation metrics. The results show a consistent increase in publications related to AI-based assessment, with articles being the primary format. Keyword analysis revealed dominant clusters centered on student perceptions and automated grading systems. This study provides an updated bibliometric landscape that guides researchers in identifying research gaps and emerging directions in AI assessment, while highlighting how AI-powered search tools can enhance systematic literature mapping.

INTRODUCTION

Speaking proficiency is a key indicator of communicative competence in second language learning, yet its assessment remains one of the most challenging areas in Malaysian primary ESL education. Although CEFR alignment has been mainstreamed into the national curriculum, classroom-based speaking assessments continue to exhibit inconsistency, particularly in rural schools where resources, training, and exposure to English are limited. Teachers often rely on impressionistic judgements or locally improvised checklists that lack standardisation, reducing scoring fairness and weakening feedback quality (Hashim & Yunus, 2020; Mohammed et al., 2021). These issues are compounded by the absence of validated analytic rubrics designed specifically for young ESL learners, resulting in assessments that inadequately capture communicative ability and offer limited instructional guidance.

RESEARCH BACKGROUND

Globally, the literature emphasises that reliable speaking assessment requires rubrics constructed from well-defined constructs and supported by empirical evidence (Fulcher, 2022). The CEFR provides calibrated descriptors for communicative performance, yet several studies across Southeast Asia highlight that classroom adaptations of CEFR descriptors are often superficial and rarely subjected to systematic validation (Butler, 2018).
Recent research further indicates that rubrics lacking psychometric testing may exhibit category disordering, inconsistent discrimination, or ambiguous descriptor interpretation, ultimately

https://doi.org/10.37134/ejoss.vol12.Sp.13.2026

compromising assessment validity (Sureeyatanapas, 2024; Shang, 2024). This underscores the need for rubrics that are not only aligned with CEFR conceptualisations but are empirically verified to function as intended in real classroom settings.

In Malaysia, the challenge is particularly pronounced in rural schools where learners consistently underperform in oral English due to limited linguistic exposure and reduced communicative opportunities (Noor et al., 2020). Assessment tools developed for urban or mixed contexts may not accurately reflect the linguistic profiles of rural learners, leading to misinterpretation of ability or inappropriate instructional decisions. However, despite ongoing interest in CEFR-informed teaching, very few studies have focused on validating speaking rubrics tailored to the rural primary context.

This gap presents a clear need for a theoretically grounded and empirically validated analytic speaking rubric that can be used confidently by teachers in rural Malaysian classrooms. The current study addresses this need by adopting a multi-stage validation design, combining expert content validation, and Classical Test Theory reliability indices. Messick's unified validity framework underpins this process, emphasising that content adequacy, scoring consistency, scale functioning, and interpretive accuracy must collectively contribute to a defensible validity argument.

Accordingly, this study has the objectives of (1) to establish the content validity of a CEFR-aligned analytic speaking rubric through expert judgement, and (2) to examine scoring consistency through inter-rater reliability and internal consistency estimates.

By focusing exclusively on rubric development and validation, this study offers both methodological and practical contributions. Methodologically, it demonstrates a systematic and transparent validation pathway rarely applied in Malaysian primary speaking assessment research. Practically, it provides teachers with a calibrated, evidence-based assessment tool capable of supporting fair scoring, diagnostic feedback, and CEFR-aligned proficiency reporting.

LITERATURE REVIEW

The CEFR provides a globally recognised framework for describing and assessing communicative competencies through calibrated and developmentally sequenced descriptors (Council of Europe, 2020). Although widely adopted, research across Southeast Asia indicates that localized development of CEFR-aligned speaking assessment instruments, particularly for young learners, remains limited and uneven (Butler, 2018). Many rubrics used in primary settings are adapted superficially without empirical calibration, leading to inconsistent judgments and reduced interpretive accuracy. North (2014) further argues that CEFR descriptors require contextual adaptation and empirical verification to ensure alignment with local linguistic realities. Hence, validated speaking rubrics must be age-appropriate, culturally relevant, and grounded in construct representations derived from CEFR descriptors.

Within language assessment, scoring rubrics serve as operational tools that translate abstract constructs into observable performance indicators. High-quality rubrics must therefore demonstrate both definitional clarity and empirical functioning (Fulcher, 2022). Recent research shows that poorly defined descriptors, ambiguous level distinctions, or untested rating categories can negatively impact scoring reliability and fairness, especially in analytic scales used with young learners (Shang, 2024). This concern highlights the need for systematic rubric validation that extends beyond superficial face agreement.

Instrument validation in speaking assessment commonly begins with content validation, where experts judge the representativeness and appropriateness of rubric indicators. The Content Validity Index (I-CVI and S-CVI) provides a widely accepted method for quantifying expert agreement and identifying descriptors that require refinement (Polit & Beck, 2006). However, content validation alone does not guarantee that a rubric function as intended during actual scoring.

Reliability evidence is also required to demonstrate scoring consistency, including inter-rater reliability indices such as Cohen's Kappa and internal consistency measures such as Cronbach's alpha (McHugh, 2012; Stemler, 2004). These indices determine whether raters interpret descriptors similarly and whether rubric domains operate cohesively as related components of a broader construct.

https://doi.org/10.37134/ejoss.vol12.Sp.13.2026

Theoretical Framework

This study is underpinned by two complimentary theoretical perspectives supporting the design, refinement and validation of analytic speaking rubrics.

According to Messick's Unified Theory of Validity (1995), validity is a single construct that integrates theoretical, empirical, and interpretive evidence across the content, construct, and consequential dimensions. According to the theory, rubric descriptions must reflect key components of oral competency, produce consistent scoring interpretations and facilitate significant instructional and evaluate decision-making in speaking assessment contexts in order to be considered legitimate. This framework directs the creation of rubrics by assessing ho well descriptors capture speaking domains that are in line with the CEFR, analysing rater interpretation and rubric usefulness and using percentile classification and proficiency ranges to prove interpretive utility. Messick therefore offers the theoretical rationale for blending construct definition, scoring interpretation and the desired outcomes of assessment use together (Tavakol & Dennick, 2011).

Classical Test Theory (CTT) is a fundamental foundation for evaluating an assessment tool's reliability. It is predicted on the idea that each score a student receives consists of both a genuine score and some mistake. This inaccuracy should be minimised by a good rubric so that ratings accurately represent a learner's speaking proficiency (DeVellis, 201). For rubric validation, CTT offers two significant indicators. First, the Internal consistency (Cronbach Alpha) demonstrates how well the many rubric domains such as coherence, pronunciation and fluency work together. A higher alpha indicates that the domains are evaluating relevant aspects of speech and that the scoring system is cohesive (Tavakol & Dennick, 2011). Next, Inter-Rater Consistency (Cohen's Kappa) is the consistency with which various raters score the same student is demonstrated. Higher Kappa values indicate that the rubric is sufficiently clear to direct scoring that raters perceive the descriptions similarly. When combined, these CTT measures offer the first level of proof that the rubric is acceptable (McHugh, 2012).

RESEARCH METHODOLOGY

This study employed a quantitative instrument development and validation design to construct and evaluate a CEFR-aligned analytic speaking rubric for upper primary ESL learners in rural Malaysia. The validation process followed established procedures in language assessment research, combining expert judgement, reliability analysis, and Rasch measurement modelling to ensure the rubric's quality and interpretive robustness.

Six TESL experts, four from Institut Pendidikan Guru and the other two are primary school teachers who have more than 20 years teaching experience were purposefully selected based on their expertise in CEFR and language assessment. Their role was to evaluate the relevance, clarity and representativeness of the draft rubric descriptors. Thirty upper primary school pupils from two rural schools participated in the pilot testing. This group provided real-world data needed to assess the rubric's reliability, scale functioning and measurement properties.

The analytic rubric was developed using CEFR-aligned descriptions suitable for young learners at the A2 level. Five dimensions of speaking performance; fluency, coherence, pronunciation, vocabulary use and interaction were found to be the main aspects. A six-point rating system was created by covering each dimension into observable indicators and categorising them into three achievement categories.
The validation process was conducted in two stages to establish the instrument's quality and reliability. For content validity, the expert panel rated the relevance of each rubric item using a 4-point Likert scale. The Item-Level Content Validity Index (I-CVI) and Scale-Level Content Validity Index (S-CVI) were calculated following Polit and Beck's (2006) guidelines to ensure that the instrument accurately measures the intended constructs.

As for the Inter-Rater Reliability, two raters independently scored the pilot participants' responses. To measure consistency between raters, Cohen's Kappa was used to assess the level of agreement, while Cronbach's alpha was calculated to evaluate the internal consistency of the rubric dimensions. This methodological approach ensures that the speaking rubric is not only theoretically grounded but also statistically validated, making it a reliable and contextually appropriate tool for assessing oral proficiency among upper primary ESL learners in rural Malaysia.

https://doi.org/10.37134/ejoss.vol12.Sp.13.2026

FINDINGS

Content Validity

The content validity of the speaking rubric was evaluated using the Item-Level Content Validity Index (I-CVI) and Scale-Level Content Validity Index (S-CVI/Ave) based on ratings from six TESL experts (see Table 1). All five rubric dimensions surpassed the recommended I-CVI threshold of 0.78 (Polit & Beck, 2006). Three domains, Fluency, Pronunciation, and Interaction achieved I-CVI = 1.00, indicating unanimous agreement among experts regarding their relevance. The other two domains, Vocabulary and Coherence, obtained I-CVI values of 0.83, also meeting the acceptable standard.

The S-CVI = 0.93 further indicates excellent scale-level validity, demonstrating strong expert consensus on the appropriateness and clarity of the rubric for assessing CEFR-A2 speaking tasks. These results confirm that the rubric items comprehensively represent the intended construct and are contextually aligned with the learning needs of upper primary ESL learners in rural Malaysia. Table 1 shows the result of content validity.

Table 1. Content Validity Index (CVI) for the Speaking Rubric

Rubric Item	I-CVI	Threshold
Fluency	1.00	≥ 0.78
Pronunciation	1.00	≥ 0.78
Vocabulary	0.83	≥ 0.78
Coherence	0.83	≥ 0.78
Interaction	1.00	≥ 0.78
S-CVI	**0.93**	**≥ 0.90**

Inter-Rater Reliability

The inter-rater reliability of the rubric was evaluated using Cohen's Kappa (κ**)** across 30 pilot responses (see Table 2). The κ values ranged from 0.61 to 0.76, demonstrating substantial agreement between the two trained raters (Landis & Koch, 1977). Among the five domains, Interaction (κ = 0.76) and Fluency (κ = 0.72) showed the strongest agreement, suggesting that raters were highly consistent in judging these aspects of speaking performance.

Although Coherence (κ = 0.61) recorded the lowest agreement, it still fell within the substantial range. This slight variability may indicate the need for more precise behavioural descriptors and exemplar scoring guides to further enhance rater consistency in this dimension.

Table 2. Inter-Rater Reliability Analysis (n = 30)

Domain	Kappa (κ)	Interpretation
Fluency	0.72	Substantial
Pronunciation	0.66	Substantial
Vocabulary	0.69	Substantial
Coherence	0.61	Substantial
Interaction	0.76	Substantial

https://doi.org/10.37134/ejoss.vol12.Sp.13.2026

Internal consistency

Analysis of internal consistency using Cronbach's Alpha yielded a coefficient of α = 0.84 (see Table 3), indicating good reliability (Tavakol & Dennick, 2011). This demonstrates that the five rubric domains measure related but distinct aspects of speaking proficiency, making the instrument robust and pedagogically sound.

Table 3. Internal Consistency Reliability for Speaking Rubric

Cronbach's Alpha	N of Items
0.84	5

DISCUSSION

This study aimed to develop and validate a CEFR-aligned analytic speaking rubric for upper primary ESL learners in rural Malaysia. The discussion synthesises evidence from content validity and Classical Test Theory (CTT) reliability analyses to build a coherent validity argument for the rubric.

First, the findings provide strong support for content validity. All rubric domains exceeded the recommended I-CVI threshold, and the high S-CVI indicates substantial expert consensus regarding the relevance, clarity, and representativeness of the descriptors. This suggests that the rubric domains fluency, pronunciation, vocabulary use, coherence, and interaction adequately operationalise key aspects of speaking proficiency expected at the CEFR A2 level. The consistency of expert judgments further indicates that the descriptors are developmentally appropriate for young learners and contextually suitable for rural ESL classrooms. In line with established validation practices, these results affirm that the rubric reflects the intended construct rather than superficial or loosely defined criteria (Polit & Beck, 2006).

Second, CTT-based reliability evidence demonstrates that the rubric functions consistently during scoring. The internal consistency coefficient (Cronbach's alpha) exceeded the commonly accepted benchmark, indicating that the five domains operate cohesively as related components of a broader speaking construct. This supports the use of an analytic scoring approach in which multiple domains contribute meaningfully to an overall proficiency interpretation (Tavakol & Dennick, 2011). Importantly, the alpha value suggests adequate homogeneity without redundancy, implying that each domain captures a distinct yet complementary aspect of oral performance.

Inter-rater reliability results further strengthen the rubric's reliability argument. Cohen's Kappa values indicated moderate to substantial agreement across domains, demonstrating that different raters were able to apply the descriptors in a largely consistent manner. Higher agreement in fluency and interaction suggests that these domains are more readily observable and clearly specified, while comparatively lower agreement in coherence reflects the interpretive complexity associated with organisational features of speech.
This pattern aligns with prior research indicating that higher-order discourse features are more challenging to judge reliably and may benefit from continued rater calibration or descriptor refinement (McHugh, 2012).

Taken together, the convergence of high content validity and acceptable reliability indices provides robust initial evidence that the rubric is both conceptually sound and scoring-stable. Following Messick's unified view of validity, these findings indicate that the rubric's construct representation and response processes are sufficiently supported for classroom use. Although the study did not extend to latent trait modelling, the combined CVI and CTT evidence establishes a defensible foundation for the rubric's application as a formative and summative assessment tool in similar contexts.

From a practical perspective, the validated rubric offers teachers a structured and transparent framework for evaluating speaking performance, reducing reliance on impressionistic judgement and supporting more consistent feedback. For rural classrooms in particular, where access to standardised assessment resources is limited, the rubric provides an evidence-based tool aligned with CEFR expectations and local learner profiles.

https://doi.org/10.37134/ejoss.vol12.Sp.13.2026

Nevertheless, as this study was conducted as a pilot with a relatively small sample, further validation with larger and more diverse cohorts is recommended. Future studies may incorporate advanced measurement modelling or longitudinal data to strengthen the rubric's generalisability and examine its sensitivity to learner development over time.

CONCLUSION

This study developed and validated a CEFR-aligned analytic speaking rubric designed for upper primary ESL learners in rural Malaysia. Drawing on expert judgment and Classical Test Theory reliability evidence, the study provides initial but robust support for the rubric's validity and reliability as a classroom-based assessment tool.

High content validity indices indicate strong expert consensus regarding the relevance, clarity, and developmental appropriateness of the rubric descriptors. This confirms that the five analytic domains fluency, pronunciation, vocabulary use, coherence, and interaction, adequately represent key components of speaking proficiency at the targeted CEFR level. In addition, internal consistency and inter-rater reliability results demonstrate that the rubric can be applied with acceptable scoring stability, supporting its use for consistent and fair evaluation of learners' oral performance.

Together, the CVI and CTT findings establish a defensible foundation for the rubric's use in similar educational contexts. The rubric offers practical value by providing teachers with a structured, transparent framework for assessing speaking skills and delivering targeted feedback, particularly in rural classrooms where standardised assessment resources are limited.

As a pilot validation study, this research has certain limitations, including a small sample size and reliance on initial reliability evidence. Future research should involve larger and more diverse learner populations and may incorporate advanced measurement approaches to further strengthen the rubric's psychometric properties and generalisability.

Overall, this study contributes to language assessment research by offering a context-responsive, CEFR-aligned speaking rubric supported by systematic validation procedures. The findings underscore the importance of empirically grounded assessment tools in enhancing the quality and fairness of speaking assessment in primary ESL education.

DECLARATION OF GENERATIVE AI

During the preparation of this work, the author(s) used ChatGPT/Perplexity/SciSpace to enhance the clarity of the writing. After using the ChatGPT/Perplexity/SciSpace, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFERENCES

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford University Press.

Butler, Y. G. (2018). English language education among young learners in East Asia: A review of current research. Language Teaching, 51(1), 1–34. https://doi.org/10.1017/S0261444817000232

Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume. Council of Europe Publishing. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2020

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage Publications.

Hashim, H., & Yunus, M. M. (2020). English language learning challenges in rural schools and strategies to improve proficiency. Creative Education, 11(6), 1045–1055. https://doi.org/10.4236/ce.2020.116078

https://doi.org/10.37134/ejoss.vol12.Sp.13.2026

Hashim, H., & Yunus, M. M. (2020). ESL teachers' perceptions on the use of speaking assessment in Malaysian primary schools. *International Journal of Learning, Teaching and Educational Research, 19*(6), 1–15. https://doi.org/10.26803/ijlter.19.6.1

Leong, L. M., & Ahmadi, S. M. (2017). An analysis of factors influencing learners' English speaking skill. International Journal of Research in English Education, 2(1), 34–41. https://doi.org/10.18869/acadpub.ijree.2.1.34

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mohammed, S. A., Rahman, M. S., & Yunus, M. M. (2021). Teachers' readiness in implementing CEFR-aligned English curriculum in rural Malaysia. Asian Journal of Education and Social Studies, 17(4), 21–34. https://doi.org/10.9734/AJESS/2021/v17i430422

Mohammed, A. A., Yunus, M. M., & Hashim, H. (2021). Challenges in assessing speaking skills among Malaysian ESL learners. *Asian Journal of University Education, 17*(3), 1–11. https://doi.org/10.24191/ajue.v17i3.14575

Mustapha, S. M., & Yahaya, A. (2013). Low English proficiency among Malaysian learners: Causes and strategies. English Language Teaching, 6(12), 155–163. https://doi.org/10.5539/elt.v6n12p155

Noor, N. M., Mohamed, A. R., & Rahman, N. A. (2020). Achieving CEFR benchmarks in Malaysian ESL primary schools: Challenges and implications. Asian EFL Journal, 24(3), 75–98.

Noor, N. M., Aman, I., Mustaffa, R., & Seong, T. K. (2020). Language proficiency and learning challenges among rural ESL learners in Malaysia. *Issues in Educational Research, 30*(2), 533–551.

North, B. (2014). Putting the Common European Framework of Reference to good use. Language Teaching, 47(2), 228–249. https://doi.org/10.1017/S0261444812000205

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29(5), 489–497. https://doi.org/10.1002/nur.20147

Shang, Y. (2024). A meta-analysis of the reliability of second language proficiency assessments. Language Assessment Quarterly, 21(2), 145–169. https://doi.org/10.1080/15434303.2024.1187649

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9(4), 1–19. https://doi.org/10.7275/96jp-xz07

Sureeyatanapas, P. (2024). Analysing marking reliability in English speaking proficiency tests: A marking consistency approach. Language Testing in Asia, 14(1), 65–79. https://doi.org/10.1186/s40468-023-00271-z

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

https://doi.org/10.37134/ejoss.vol12.Sp.13.2026