Al-Ishlah: Jurnal Pendidikan
Vol.18, 1 (March, 2026), pp. 504-520
ISSN: 2087-9490 EISSN: 2597-940X, DOI: 10.35445/alishlah.v18i1.9642

Psychometric Evidence for Using FSI Speaking Ratings in Indonesian Primary EFL Classrooms: Content Validity and Inter-Rater Reliability

Euis Yanah Mulyanah[1] , Yudi Juniardi[2], Lukman Nulhakim[3]

[1] Universitas Sultan Ageng Tirtayasa, Serang, Indonesia; 7782240002@student.untirta.ac.id
[2] Universitas Sultan Ageng Tirtayasa, Serang, Indonesia; yudi.juniardi@untirta.ac.id
[3] Universitas Sultan Ageng Tirtayasa, Serang, Indonesia; lukman.nulhakim@untirta.ac.id

ARTICLE INFO

Keywords:

psychometric validation;
reliability between appraisers;
speaking skills assessment;
foreign service institute (FSI);
EFL primary school learners

Article history:

Received 2026-02-18
Revised 2026-03-01
Accepted 2026-03-30

ABSTRACT

Reliable and valid speaking assessment is crucial for accurately interpreting young learners' communicative competence in English as a Foreign Language (EFL). Although the Foreign Service Institute (FSI) Speaking Ratings are widely used in adult contexts, empirical evidence supporting their adaptation for primary school learners, particularly in Indonesia, remains limited. This study employed a quantitative psychometric validation design to examine the content validity and inter-rater reliability of an adapted FSI scale. Six expert validators (two media, two language, and two material experts), two trained raters, and 30 Grade V students from a public primary school participated. The scale was contextually modified to align with young learners' characteristics while retaining its five domains. Students performed a 2–3 minute monologue based on visual prompts, which was video-recorded and independently scored. Content validity was assessed using Aiken's V, while inter-rater reliability was analysed using a two-way random-effects Intraclass Correlation Coefficient (ICC) with absolute agreement Aiken's V coefficients ranged from 0.50 to 1.00, with a mean of 0.87 across 54 indicators, indicating strong content validity. The ICC results demonstrated consistent scoring between raters, suggesting satisfactory inter-rater reliability. The findings provide initial psychometric support for the adapted FSI Speaking Ratings in primary EFL contexts, enhancing assessment objectivity and standardization. However, limitations include a small sample size, limited number of raters, single-site data, and the absence of construct validity analysis. Future studies should address these constraints to strengthen generalizability and validation.

Corresponding Author:
Euis Yanah Mulyanah
Universitas Sultan Ageng Tirtayasa, Serang, Indonesia; 7782240002@student.untirta.ac.id

1. INTRODUCTION

The assessment of speaking skills in Indonesian primary EFL classrooms remains both essential and insufficiently developed, particularly in terms of ensuring fairness, consistency, and accountability in evaluating learners' communicative competence. Existing studies suggest that conventional

assessment practices, often shaped by cultural expectations and limited use of structured evaluation frameworks, may constrain the accurate measurement of students' oral proficiency and hinder the systematic development of communicative skills (Halim et al., 2025; Widiastuti, 2025). To promote greater objectivity and reliability, speaking assessments should be contextually grounded and guided by clearly articulated performance criteria, as emphasized by Brown and Abeywickrama (2020), a principle further reinforced in recent Indonesian EFL research (Widiastuti, 2025). Moreover, the implementation of formative assessment practices has been shown to strengthen students' speaking development by providing structured feedback, encouraging self-reflection, and supporting continuous performance improvement (Maulida et al., 2025). Nevertheless, persistent challenges—including teacher-centered instructional traditions, limited opportunities for authentic language exposure, and insufficient professional training in assessment literacy continue to impede the effectiveness of classroom-based speaking evaluation (Arsyad & Suadiyatno, 2024; Halim et al., 2025). Addressing these systemic concerns through differentiated assessment strategies and sustained teacher professional development initiatives is, therefore, essential to fostering a more equitable, transparent, and pedagogically sound speaking assessment environment in Indonesian primary EFL contexts (Arsyad & Suadiyatno, 2024).

The validation of speaking assessment instruments is fundamentally grounded in Kane's argument-based validity framework, which conceptualizes validity as the strength of an interpretive argument linking observed performance to score meaning and subsequent decision-making. This framework emphasizes the systematic integration of multiple sources of validity evidence, particularly content validity and scoring reliability, to ensure that assessment outcomes accurately represent the intended construct and are applied consistently across raters (Lauwaert, 2023; Raymond et al., 2025). Content validity is especially critical in primary EFL contexts, as it ensures that assessment tasks, descriptors, and criteria are developmentally appropriate and aligned with young learners' linguistic and cognitive characteristics. At the same time, scoring reliability, particularly inter-rater reliability, ensures that scores are applied in a stable and consistent manner, thereby enhancing the dependability and interpretability of assessment results (Lane & Marion, 2025). Furthermore, this argument-based perspective aligns closely with Messick's unified validity framework, which highlights the importance of coherent construct representation and score consistency as essential foundations for valid and defensible assessment practices across diverse educational contexts (Raymond et al., 2025).

The Foreign Service Institute (FSI) Speaking Ratings scale, while theoretically robust for assessing speaking proficiency, presents challenges when applied to primary school learners due to developmental and contextual factors. The scale's multidimensional assessment of pronunciation, grammar, vocabulary, fluency, and comprehension aligns with contemporary understandings of communicative competence, which emphasize not only linguistic accuracy but also cultural awareness and adaptability in real-world contexts (Sultana Shaik, 2024). Moreover, effective assessment methods for young learners should incorporate diverse approaches, such as performance-based assessments and task-based language teaching, to address their unique communicative needs (Saptiany & Prabowo, 2024). Additionally, the integration of both linguistic and functional dimensions in assessment is crucial, as it ensures a comprehensive evaluation of speaking abilities that reflects the learners' practical communication skills (Заирова, 2023). Therefore, while the FSI scale provides a solid foundation, its direct application necessitates careful adaptation to suit the developmental stages and contextual realities of primary school students. Young EFL learners differ substantially from adult learners in terms of cognitive maturity, linguistic development, attentional capacity, and communicative experience, all of which have important implications for assessment performance and score interpretation. Research has shown that children benefit most from interactive, engaging, and developmentally appropriate learning environments that foster emotional involvement and facilitate language acquisition, in contrast to adult learners who often rely more heavily on explicit instruction and external learning resources (Estrada Ramos & Hernández Alipi, 2025). In this regard, the original Foreign Service Institute (FSI) speaking descriptors, which were designed primarily for adult learners in formal or professional contexts, may not fully capture the simpler communicative demands,

developmental characteristics, and pedagogical objectives associated with young learners' language use (Quesada Pacheco, 2023). Empirical validation studies of speaking performance assessments further suggest that existing rating scales may not adequately reflect the emerging and variable nature of young learners' communicative abilities, highlighting the need for careful adaptation to ensure developmental appropriateness and accurate construct representation (Joo & Lee, 2022). Therefore, the modification and empirical validation of speaking assessment instruments are essential to ensure that score interpretations are meaningful, developmentally appropriate, and supportive of young learners' linguistic and cognitive growth within child-centered educational environments (Nguyen, 2021).

The application of the FSI Speaking Ratings scale in primary EFL contexts, particularly in Indonesia, reveals significant gaps in the literature that warrant attention. Existing validation studies primarily focus on adult learners and higher education, neglecting the unique needs of primary school learners, which raises concerns about the scale's validity and reliability in this demographic (Halim et al., 2025; Widiastuti, 2025). Furthermore, the inter-rater reliability of FSI-based scoring in Indonesian classrooms remains underexplored, particularly given the variability in rater training and experience (Arsyad & Suadiyatno, 2024). The suitability of FSI descriptors for young learners is also inadequately examined, especially regarding their developmental language stages and the context of classroom assessments (Putri & Sya, 2023). These gaps are critical, as they challenge the argument-based validity framework, which emphasizes the importance of content relevance and scoring consistency for valid score interpretation (Gultom et al., 2024). Addressing these issues is essential for establishing the FSI scale's effectiveness in assessing speaking proficiency among Indonesian primary EFL learners.

The investigation into the psychometric properties of the adapted FSI Speaking Ratings scale for Indonesian primary EFL learners is crucial for establishing its validity and reliability in assessment contexts. Expert judgments are essential for supporting content validity, as they ensure that the descriptors align with the specific linguistic and cultural needs of the learners (Jung Youn, 2023). Furthermore, inter-rater reliability is vital, as consistent scoring among independent raters indicates the scale's robustness and fairness in evaluating speaking proficiency ((Đorđević, 2025). The empirical support for the scale's contextual appropriateness can be bolstered by examining its alignment with established constructs of communicative adequacy, which emphasizes the importance of complexity, accuracy, fluency, and pronunciation in speaking assessments (Chen, 2025). By addressing these aspects, the study aims to contribute to the development of reliable assessment practices that enhance the learning experience for young language learners in Indonesia. (Neiriz, 2023).

To address these gaps, this study investigates the psychometric properties of an adapted version of the FSI Speaking Ratings scale in an Indonesian primary EFL context. Specifically, this study seeks to answer the following research questions: (1) To what extent do expert judgments support the content validity of the adapted FSI Speaking Ratings descriptors for Indonesian primary EFL learners? (2) To what extent do independent raters demonstrate consistent scoring when using the adapted FSI Speaking Ratings scale, as evidenced by inter-rater reliability estimates? and (3) To what extent does the adapted FSI Speaking Ratings scale provide empirical support for its use as a contextually appropriate and psychometrically defensible speaking assessment tool for Indonesian primary EFL classrooms? By addressing these questions, this study aims to provide empirical validity evidence to support the interpretation and use of FSI-based speaking assessment scores in primary EFL education, thereby contributing to the development of fair, reliable, and theoretically grounded assessment practices for young language learners.

2. METHODS

The study conducted at SDN Karet II in Tangerang Regency, Indonesia, aimed to validate a speaking assessment scale for primary school EFL learners, emphasizing the importance of instrument validation in ensuring accurate and consistent score interpretations aligned with the intended construct (Vo, 2017). Ethical considerations were paramount, with procedures including parental consent and data anonymization, adhering to established standards for protecting child participants (Ölmezer-

Öztürk & Aydin, 2018). The validation process is critical, as highlighted by Huang et al. (2021), who noted the lack of validation research in speaking assessments for EFL learners, underscoring the need for reliable instruments in educational contexts (Huang et al., 2021). Furthermore, the study aligns with broader trends in K-12 language assessments, which often grapple with the validity-reliability paradox, necessitating rigorous validation processes to support effective language support programs (Sinclair & Lau, 2018).

The participants consisted of 30 fourth-grade students aged 9–10 years, including 21 females and 9 males, all of whom demonstrated elementary-level English proficiency and limited exposure to English primarily through formal classroom instruction conducted twice per week. This sample size was considered appropriate for a pilot validation study, consistent with established practices in the initial phase of instrument development aimed at evaluating psychometric properties prior to large-scale implementation.

The study in question utilized the Foreign Service Institute (FSI) Speaking Ratings scale, adapted into an analytic scoring rubric with six proficiency levels and five core dimensions: pronunciation, grammar, vocabulary, fluency, and comprehension. This approach aligns with the findings of Gao and Sun, who emphasize the importance of considering different fluency dimensions in L2 speaking assessments, particularly in distinguishing between monologic and dialogic tasks (Gao & Sun, 2025). The analytic scoring method enhances diagnostic precision by allowing for a more detailed evaluation of communicative competence compared to holistic scoring approaches, as supported by the work of Zhang et al., who highlight the role of expert judgment in establishing content validity through a comprehensive set of indicators (X. Zhang & Lu, 2025). The integration of pronunciation-focused discussions, as explored by Mister, further underscores the significance of pronunciation in communicative competence, which is often overlooked but crucial for intelligibility and confidence in real-world contexts (Mister, 2025). Additionally, the study's focus on fluency is echoed in the research by Suzuki and Kormos, who investigate how task demands affect fluency, revealing that different tasks require varying levels of conceptualization and formulation, impacting fluency measures such as speed and pause-frequency (Suzuki & Kormos, 2025). The use of technology in language assessment, as discussed by Liao, also plays a role in enhancing speaking proficiency and self-efficacy, suggesting that self-assessment can lead to greater gains in accuracy and linguistic complexity (Liao, 2025). Overall, the study's methodology and focus on detailed, dimension-specific evaluation are well-supported by the broader literature on language assessment and fluency, highlighting the importance of a nuanced approach to evaluating L2 speaking skills.

The speaking task consisted of an individual monologue lasting approximately 2–3 minutes, in which students were instructed to describe their family members. Monologue tasks are considered effective for eliciting authentic spoken language production and ensuring consistency in performance-based language assessment. All student performances were video-recorded and independently evaluated by two raters—a primary school English teacher and a university lecturer in English education—following structured briefing, calibration, and practice sessions using benchmark video exemplars. Rater training and calibration procedures have been shown to significantly enhance scoring consistency and reliability in performance-based language assessment.

Content validity is a critical aspect of psychometric assessment, ensuring that a measure accurately represents the construct it aims to assess. Aiken's V coefficient is a widely used method for quantifying expert agreement on item relevance, as demonstrated in the validation of instruments for evaluating perceptions of renewable energy and energy sustainability, where Aiken's V values exceeded 0.75, indicating strong content validity (Acosta-Banda et al., 2021). The use of expert judgment is a common approach, as seen in the validation of an instrument for measuring actions mediated by technology in educational settings, where expert evaluations confirmed high content validity and reliability through various indices, including Cronbach's alpha (Borja & Navarro, 2023). Formal content validity analysis (FCVA) and its Bayesian extension (B-FCVA) offer advanced methodologies for evaluating content validity by combining traditional methods with Bayesian procedures to enhance the accuracy of interrater agreement indices (Spoto, 2025). The integration of Item Response Theory (IRT)

in content validity assessment further refines the evaluation process by estimating discrimination and threshold parameters, thereby improving the alignment of items with the intended constructs (Kreitchmann et al., 2024). In the context of early numeracy measures, a systematic review using the COSMIN framework highlighted the importance of comprehensive content validity evaluation, although no measures were recommended due to insufficient high-quality evidence (Speyer et al., 2024). Additionally, the use of Large Language Models (LLMs) in conjunction with human expertise has shown potential in enhancing the content validity assessment of psychometric instruments, particularly in aligning items with constructs in personality tests (Milano et al., 2026). These diverse methodologies underscore the multifaceted nature of content validity assessment, emphasizing the need for robust, multi-method approaches to ensure the accuracy and reliability of psychometric instruments across various domains.

3. FINDING AND DISCUSSION

3.1 Findings

Table 1. Raw Inter-Rater Scores of FSI Speaking Dimensions (Pilot Test)

Student	Pron_R1	Pron_R2	Gram_R1	Gram_R2	Voc_R1	Voc_R2	Flu_R1	Flu_R2	Comp_R1	Comp_R2
S1PT	3	3	30	36	20	12	10	10	19	19
S2PT	2	3	24	30	12	16	8	6	12	12
S3PT	4	4	30	36	20	20	10	12	23	23
S4PT	0	0	6	6	4	4	2	2	4	4
S5PT	1	0	12	6	4	4	4	2	8	4
S6PT	4	4	30	36	20	24	10	12	23	23
S7PT	1	1	12	24	8	16	4	8	8	15
S8PT	1	1	12	12	4	8	2	4	4	8
S9PT	1	1	6	12	4	8	2	4	4	8
S10PT	3	4	30	36	20	24	10	12	19	23
S11PT	3	4	30	36	20	24	10	12	19	23
S12PT	3	2	30	24	20	16	10	8	19	15
S13PT	3	3	30	30	20	24	10	12	19	23
S14PT	2	2	18	18	12	12	6	6	12	12
S15PT	4	4	30	36	20	24	10	12	19	23
S16PT	1	1	12	12	8	8	4	4	8	8
S17PT	3	4	30	36	20	24	10	12	19	23
S18PT	4	4	30	36	20	24	10	12	19	23
S19PT	3	2	30	24	20	16	10	8	19	15
S20PT	3	2	30	24	20	16	10	8	19	15
S21PT	4	4	36	36	20	24	10	12	23	23
S22PT	3	2	30	18	20	12	10	6	19	12
S23PT	2	3	24	30	16	20	6	10	12	19
S24PT	3	2	30	18	20	12	10	6	19	12
S25PT	4	4	36	36	24	24	12	12	23	23
S26PT	3	3	30	30	20	20	10	10	19	19
S27PT	3	2	30	12	20	8	10	4	19	8
S28PT	2	3	24	30	12	20	6	10	12	19
S29PT	4	3	36	30	24	20	10	10	23	19
S30PT	2	2	24	18	16	12	6	6	12	12

Table 1 presents the raw inter-rater scores obtained from two independent raters who evaluated students' speaking performance using the adapted analytic FSI Speaking Rating scale. The table displays individual student scores across five key dimensions of speaking proficiency, namely pronunciation, grammar, vocabulary, fluency, and comprehension. These raw scores provide the empirical basis for subsequent psychometric analyses, including the estimation of inter-rater reliability using the Intraclass Correlation Coefficient (ICC) and the calculation of the Standard Error of Measurement (SEM). Reporting the raw score matrix is essential to ensure transparency in the scoring process and to demonstrate the consistency and variability of ratings across raters and speaking dimensions, thereby supporting the validity argument for the use of the instrument in the context of primary school EFL assessment (Istihari et al., 2025).

Table 2. Item-Level Content Validity Results of the EUIS-Based Adapted FSI Speaking Assessment Instrument Based on Expert Judgment Using Aiken's V

Dimension	Aspect	Indicator Number	Indicator	Aiken V	Interpretation
E-Visual English Instructions (EUIS) Learning Media	Usability	1	Ease of learning to use the media by primary school students	0.875	Valid
		2	Ease of operating features and main menu	0.75	Revise
		3	Ease of use without intensive teacher assistance	0.75	Revise
	Visual Design (Simple UI/UX)	4	Simplicity and readability of the visual display	0.875	Valid
		5	Design suitability for primary school students	1	Valid
		6	Clarity of icons, colours, and layout	1	Valid
	Instructional Clarity	7	Clarity of instruction wording	0.75	Revise
		8	Ease of understanding instructions for primary school students	0.875	Valid
		9	Alignment of instructions with learning steps	0.875	Valid
	Application Navigation	10	Ease of navigation between menus	0.75	Revise
		11	Consistency and clarity of navigation icons	0.875	Valid
		12	Minimal confusion when using the application	0.5	Revise
	Implementation Feasibility in Primary School Context	13	Suitability of media use with primary classroom conditions	0.625	Revise
		14	Ease of implementing media in learning activities	0.625	Revise
		15	Integration of media with teacher and student roles	0.75	Revise
Activity Based Learning	ABL Activities	16	EUIS provides features that support student learning activities	0.75	Revise

(ABL dalam konteks media)	Facilitated by EUIS	17	ABL-based activity steps are clearly presented in EUIS	0.75	Revise
		18	EUIS facilitates students and teachers in conducting learning activities	0.875	Valid
	Clarity of Activity Procedure Presentation	19	Learning activity steps are presented sequentially and systematically	0.625	Revise
		20	Each activity step is easy for primary school students to understand	0.625	Revise
		21	Presentation of activity steps in EUIS facilitates learning implementation	0.875	Valid
	Media Support for Student Learning Activities	22	EUIS media encourages student active participation in learning activities	0.75	Revise
		23	EUIS features help students complete learning activities independently	0.75	Revise
		24	EUIS media supports student engagement during the learning process	1	Valid
Language	English Language Accuracy	1	Accuracy of grammar	0.875	Valid
		2	Accuracy of vocabulary and expressions	0.875	Valid
		3	Accuracy of simple sentence structure	1	Valid
	Appropriateness of Language Level for Primary School Students	4	Language difficulty level appropriate to students' age and grade	0.875	Valid
		5	Simple and communicative sentences	0.875	Valid
		6	Does not create language misconceptions	0.75	Revise
	Clarity and Comprehensibility of Language	7	English instructions are easy to understand	1	Valid
		8	Not ambiguous or open to multiple interpretations	1	Valid
		9	Consistency of language terminology	0.875	Valid
	Alignment of Language with Speaking Objectives	10	Language encourages students to speak actively	0.75	Revise
		11	Supports simple oral communication	1	Valid
		12	Relevant to speaking tasks	1	Valid
	Alignment of Language with Children's Context	13	Examples are close to students' real-life context	1	Valid
		14	Language appropriate to children's context	1	Valid
		15	Does not contain abstract or advanced terminology	1	Valid

Material	Appropriateness of Test Format	1	Performance test format aligns with speaking assessment objectives for Grade V students	1	Valid
		2	Oral interview (OPI) format allows students to speak naturally	1	Valid
		3	Test format aligns with primary school English learning characteristics	1	Valid
	Clarity of Task Instructions	4	Task instructions use simple and understandable language	1	Valid
		5	Instructions do not cause multiple interpretations	0.875	Valid
		6	Instructions support smooth implementation of the speaking test	0.75	Revise
	Alignment of Rubric with Primary Students' Development	7	Rubric aligns with students' cognitive and language development level	1	Valid
		8	Rubric aspects are relevant to Grade V students' speaking ability	1	Valid
		9	Rubric difficulty level matches average primary school student ability	0.875	Valid
	Clarity of Descriptors at Each Score Level	10	Descriptors for each score level are clearly and specifically written	1	Valid
		11	Differences between score levels are clearly distinguishable	1	Valid
		12	Descriptors support objective and consistent assessment	0.875	Valid
	Suitability of FSI Rating for Primary School Context	13	Adaptation of FSI Rating aligns with primary school student context	0.875	Valid
		14	FSI scale is relevant to primary school English learning	1	Valid
		15	FSI Rating supports fair and accurate speaking assessment	0.875	Valid

Table 2 presents the item-level content validity results of the EUIS-based adapted FSI speaking assessment instrument based on expert evaluation using Aiken's V coefficient. The findings indicate varying levels of content validity across indicators, with Aiken's V values ranging from 0.167 to 1.292. Several indicators related to usability, language accuracy, and test format—such as ease of learning to use the media (V = 0.292), accuracy of grammar (V = 0.292), and alignment of the performance test format with speaking assessment objectives (V = 0.167)—were classified as "Needs revision," suggesting insufficient agreement among experts regarding their relevance and clarity. In contrast, many indicators related to instructional design, student engagement, and contextual appropriateness demonstrated strong validity evidence, including integration of media with teacher and student roles (V = 0.833), presentation of activity steps (V = 0.958–1.125), and support for student engagement (V = 1.292), which were categorized as "Excellent content validity." Additionally, several indicators related

to communication support and contextual relevance, such as supporting simple oral communication (V = 0.750) and alignment with primary school learning context (V = 0.708), showed "Good content validity." Overall, these results provide initial evidence that the EUIS-based adapted FSI speaking instrument demonstrates acceptable to excellent content validity in key instructional and communicative aspects, although revisions are necessary for indicators related to usability, linguistic clarity, and rubric specificity to ensure stronger construct representation and suitability for primary school EFL learners (Rima et al., 2025).

Table 3. Item-Level Aiken's V Summary of the EUIS Media and Adapted FSI Speaking Assessment Instrument Based on Expert Judgment

Indicator	Validator Mean	Aiken's V	Interpretation
Ease of learning to use the media by primary school students	3.333	0.292	Needs revision
Ease of operating features and main menu	3.333	0.292	Needs revision
Ease of use without intensive teacher assistance	3.667	0.333	Needs revision
Simplicity and readability of the visual display	4.333	0,417	Needs revision
Design suitability for primary school students	5	0.5	Needs revision
Clarity of icons, colors, and layout	5.333	0.542	Needs revision
Clarity of instruction wording	5	0.5	Needs revision
Ease of understanding instructions for primary school students	5.667	0.583	Needs revision
Alignment of instructions with learning steps	6	0.625	Acceptable content validity
Ease of navigation between menus	6	0.625	Acceptable content validity
Consistency and clarity of navigation icons	6.667	0.708	Good content validity
Minimal confusion when using the application	6	0.625	Acceptable content validity
Suitability of media use with primary classroom conditions	6.667	0.708	Good content validity
Ease of implementing media in learning activities	7	0.75	Good content validity
Integration of media with teacher and student roles	7.667	0.833	Excellent content validity
EUIS provides features that support student learning activities	8	0.875	Excellent content validity
ABL-based activity steps are clearly presented in EUIS	8.333	0.917	Excellent content validity
EUIS facilitates students and teachers in conducting learning activities	9	1	Excellent content validity
Learning activity steps are presented sequentially and systematically	8.667	0.958	Excellent content validity
Each activity step is easy for primary school students to understand	9	1	Excellent content validity
Presentation of activity steps in EUIS facilitates learning implementation	10	1.125	Excellent content validity
EUIS media encourages student active participation in learning activities	10	1.125	Excellent content validity
EUIS features help students complete learning activities independently	10.333	1.167	Excellent content validity
EUIS media supports student engagement during the learning process	11.333	1.292	Excellent content validity

Accuracy of grammar	3.333	0.292	Needs revision
Accuracy of vocabulary and expressions	3.667	0.333	Needs revision
Accuracy of simple sentence structure	4.333	0.417	Needs revision
Language difficulty level appropriate to students' age and grade	4.333	0.417	Needs revision
Simple and communicative sentences	4.667	0.458	Needs revision
Does not create language misconceptions	4.667	0.458	Needs revision
English instructions are easy to understand	5.667	0.583	Needs revision
Not ambiguous or open to multiple interpretations	6	0.625	Acceptable content validity
Consistency of language terminology	6	0.625	Acceptable content validity
Language encourages students to speak actively	6	0.625	Acceptable content validity
Supports simple oral communication	7	0.75	Good content validity
Relevant to speaking tasks	7.333	0.792	Good content validity
Examples are close to students' real-life context	7.667	0.833	Excellent content validity
Language appropriate to children's context	8	0.875	Excellent content validity
Does not contain abstract or advanced terminology	8.333	0.917	Excellent content validity
Performance test format aligns with speaking assessment objectives for Grade V students	3	0.167	Needs revision
Oral interview (OPI) format allows students to speak naturally	3.5	0.208	Needs revision
Test format aligns with primary school English learning characteristics	4	0.25	Needs revision
Task instructions use simple and understandable language	4.5	0.292	Needs revision
Instructions do not cause multiple interpretations	4.5	0,292	Needs revision
Instructions support smooth implementation of the speaking test	4.5	0.292	Needs revision
Rubric aligns with students' cognitive and language development level	6	0.417	Needs revision
Rubric aspects are relevant to Grade V students' speaking ability	6.5	0.458	Needs revision
Rubric difficulty level matches average primary school student ability	6.5	0.458	Needs revision
Descriptors for each score level are clearly and specifically written	7.5	0.542	Needs revision
Differences between score levels are clearly distinguishable	8	0.583	Needs revision
Descriptors support objective and consistent assessment	8	0.583	Needs revision
Adaptation of FSI Rating aligns with primary school student context	8.5	0.625	Acceptable content validity
FSI scale is relevant to primary school English learning	9.5	0.708	Good content validity
FSI Rating supports fair and accurate speaking assessment	9.5	0.708	Good content validity

Table 3 presents the item-level content validity analysis of the EUIS-based adapted FSI speaking assessment instrument based on expert evaluation using Aiken's V coefficient, which is widely

recognized as a robust method for quantifying expert agreement regarding item relevance and construct representation in educational assessment development (Y. Zhang & Zhang, 2024). The results show substantial variability in content validity across indicators, with Aiken's V values ranging from 0.167 to 1.292, indicating that while several indicators related to usability, linguistic accuracy, and rubric clarity require revision, other indicators demonstrate acceptable to excellent levels of construct representation and pedagogical alignment. Notably, indicators related to instructional sequencing, student engagement, and contextual appropriateness—such as structured activity presentation (V = 0.958–1.125), support for independent learning (V = 1.167), and student engagement facilitation (V = 1.292)—demonstrated excellent content validity, suggesting strong alignment with theoretical principles of communicative language teaching and learner-centered digital pedagogy. In contrast, several indicators related to linguistic clarity, task format, and rubric descriptor specificity demonstrated lower Aiken's V values (V = 0.167–0.583), indicating the need for targeted revision to improve clarity, developmental appropriateness, and scoring interpretability, which is essential for strengthening the validity argument in performance-based language assessment instruments. Overall, these findings provide initial empirical evidence supporting the content validity of the adapted FSI-based EUIS instrument, while also identifying specific components requiring refinement to ensure accurate, reliable, and developmentally appropriate assessment of primary school EFL learners' speaking proficiency (Arafah et al., 2023).

Table 4. Scale-Level Aiken's V Content Validity Summary by Dimension for the EUIS Media and Adapted FSI Speaking Assessment Instrument

Dimension	Mean Aiken's V	Interpretation
ABL Activities Facilitated by EUIS	0.792	Good content validity
Media Support for Student Learning Activities	0.833	Excellent content validity
Clarity of Descriptors at Each Score Level	1	Excellent content validity
Instructional Clarity	0.833	Excellent content validity
Clarity of Task Instructions	1	Excellent content validity
Language Clarity and Comprehensibility	0.958	Excellent content validity
Clarity of Activity Procedure Presentation	0.708	Good content validity
Suitability of FSI Rating Scale for Primary School	0.875	Excellent content validity
Usability	0.792	Good content validity
Language Appropriateness for Child Context	1	Excellent content validity
Language Appropriateness for Speaking Objectives	0.917	Excellent content validity
Test Format Appropriateness	1	Excellent content validity
Rubric Appropriateness for Primary School Learner Development	1	Excellent content validity
Language Level Appropriateness for Primary School Learners	0.833	Excellent content validity
English Language Accuracy	0.917	Excellent content validity
Implementation Feasibility at the Primary School Level	0.667	Acceptable content validity
Application Navigation	0.708	Good content validity
Visual Design (Simple UI/UX)	0.958	Excellent content validity

Table 4 presents the dimension-level content validity results of the EUIS media and adapted FSI speaking assessment instrument based on expert judgment using Aiken's V coefficient, which is widely recognized as a robust method for establishing the degree of expert agreement and construct

representation in educational assessment instruments. The findings indicate that most dimensions demonstrated strong content validity, with mean Aiken's V values ranging from 0.667 to 1.000, suggesting that key components such as clarity of descriptors (V = 1.000), task instruction clarity (V = 1.000), rubric appropriateness (V = 1.000), and language appropriateness for child context (V = 1.000) achieved excellent levels of expert agreement and construct alignment. Several dimensions related to instructional support, usability, and navigation demonstrated good content validity (V = 0.708–0.833), indicating adequate alignment with pedagogical and usability principles necessary for effective digital language learning tools in primary education contexts. Although implementation feasibility at the primary school level showed slightly lower validity (V = 0.667), it remained within the acceptable range, suggesting that the instrument is generally suitable for classroom implementation with minor refinements to enhance contextual adaptability and usability. Overall, these results provide strong preliminary evidence supporting the content validity of the EUIS-based adapted FSI speaking assessment instrument at the dimensional level, reinforcing its theoretical alignment, instructional relevance, and suitability for assessing speaking proficiency among primary school EFL learners (Kaharuddin et al., 2023).

Table 5. Inter-Rater Reliability Analysis of the Adapted FSI Speaking Assessment Using ICC (Two-Way Random Effects, Absolute Agreement)

Skill	ICC(2,1) Single Measure	ICC(2,k) Average Measure	MSR (Rows)	MSC (Columns)	MSE (Error)	Interpretation
Pronunciation	0.835	0.9101	2.6	0.0667	0.2391	Good
Grammar	0.7344	0.8469	153.6207	0.6	24.1862	Moderate
Vocabulary	0.7034	0.8259	70.8414	1.0667	12.6529	Moderate
Fluency	0.7043	0.8265	17.3356	1.6667	3.046	Moderate
Comprehension	0.75	0.8572	67.2736	1.0667	9.8598	Good

Table 5 presents the inter-rater reliability results of the adapted FSI speaking assessment instrument using the Intraclass Correlation Coefficient (ICC) based on a two-way random effects model with absolute agreement, which is widely recommended for performance-based language assessment involving multiple raters. The findings indicate that single-measure ICC(2,1) values ranged from 0.703 to 0.835, while average-measure ICC(2,k) values ranged from 0.826 to 0.910, demonstrating moderate to good levels of inter-rater agreement across pronunciation, grammar, vocabulary, fluency, and comprehension. Pronunciation (ICC = 0.835) and comprehension (ICC = 0.750) demonstrated good reliability, whereas grammar (ICC = 0.734), vocabulary (ICC = 0.703), and fluency (ICC = 0.704) fell within the moderate reliability range, reflecting the inherent complexity of evaluating linguistic accuracy and expressive performance in young learners. The reported mean square values (MSR, MSC, and MSE) further indicate that variability in scores was primarily attributable to differences in student performance rather than systematic rater disagreement, thereby supporting the stability of score interpretation. Overall, these reliability estimates provide empirical evidence that the adapted FSI-based analytic speaking instrument yields sufficiently consistent scoring across raters, contributing to the broader validity argument for its use in primary school EFL assessment contexts (Mulyanah, Juniardi, et al., 2025a).

Table 6. Inter-Rater Reliability and Standard Error of Measurement (SEM) by Dimension Using ICC (Two-Way Random Effects, Absolute Agreement)

Dimension	ICC(2,1) Absolute Agreement	SD (Dimension Mean Score)	Standard Error of Measurement (SEM)	Reliability Interpretation
Pronunciation	0.835	1.1402	0.4632	Good
Grammar	0.7344	8.7642	4.5163	Moderate
Vocabulary	0.7034	5.9515	3.2411	Moderate
Fluency	0.7043	2.9441	1.601	Moderate
Comprehension	0.75	5.7997	2.8996	Good

Table 6 presents the dimension-level inter-rater reliability and measurement precision of the adapted FSI speaking assessment instrument using the Intraclass Correlation Coefficient (ICC) and Standard Error of Measurement (SEM), which are essential indicators for evaluating score consistency and measurement accuracy in performance-based language assessment. The ICC(2,1) absolute agreement values ranged from 0.703 to 0.835, indicating moderate to good inter-rater reliability across speaking dimensions, with pronunciation (ICC = 0.835) and comprehension (ICC = 0.750) demonstrating higher scoring consistency compared to grammar, vocabulary, and fluency. The Standard Error of Measurement (SEM) values ranged from 0.463 to 4.516, reflecting acceptable levels of measurement precision, with lower SEM values indicating greater score stability and more accurate estimation of students' true speaking ability. The relatively higher SEM observed in grammar and vocabulary dimensions suggests greater variability in linguistic performance, which is consistent with the developmental and cognitively demanding nature of these language components in young EFL learners. Overall, these findings provide empirical evidence that the adapted FSI-based analytic speaking assessment demonstrates acceptable reliability and measurement precision, thereby supporting its use for classroom-based evaluation and contributing to the broader validity argument for primary school speaking assessment (Mulyanah, Juniardi, et al., 2025).

3.2 Discussion

The findings of this study provide initial empirical evidence supporting the validity and reliability of the adapted FSI speaking assessment instrument for evaluating primary school EFL learners' speaking proficiency. The multidimensional raw score distributions across pronunciation, grammar, vocabulary, fluency, and comprehension demonstrate that the instrument effectively captures meaningful variability in students' communicative competence. Content validity analysis using Aiken's V showed strong expert agreement, with dimension-level coefficients ranging from 0.667 to 1.000, confirming that the instrument adequately represents the intended construct and aligns with pedagogical and developmental appropriateness. Furthermore, inter-rater reliability analysis revealed moderate to good consistency between raters, with ICC(2,1) values ranging from 0.703 to 0.835 and ICC(2,k) values ranging from 0.826 to 0.910, indicating stable and reproducible scoring across evaluators. Measurement precision was also supported by SEM values ranging from 0.463 to 4.516, suggesting that observed scores provide reasonably accurate estimates of students' true speaking ability. Taken together, these findings support the provisional use of the adapted FSI speaking instrument for classroom-based speaking assessment while highlighting the importance of continued validation efforts involving larger samples, additional raters, and expanded task conditions to strengthen the overall validity argument and generalizability of score interpretation.

The findings of this study provide initial empirical support for the validity argument of the adapted FSI speaking assessment instrument as a meaningful and reliable measure of primary school students' speaking proficiency. The alignment of the instrument's five dimensions—pronunciation, grammar, vocabulary, fluency, and comprehension—with established models of communicative competence supports strong content representation, while inter-rater reliability results (ICC range =

0.703–0.835) and acceptable measurement precision (SEM range = 0.463–4.516) demonstrate that the scoring process is sufficiently consistent and capable of producing stable estimates of students' true speaking ability. These results indicate that score variability primarily reflects genuine differences in learner performance rather than measurement error, thereby supporting the interpretive and evaluation inferences within the validity framework. However, given the pilot nature of the study, including a limited sample size, two raters, and a restricted task context, the instrument should be considered provisionally valid for classroom-based assessment, and further validation research involving larger samples, additional raters, and broader task conditions is recommended to strengthen the generalizability and overall validity of score interpretation (Mulyanah, Arwen, et al., 2025; Nulhakim et al., 2019).

4. CONCLUSION

This study provides initial evidence of content validity and inter-rater reliability for the adapted FSI Speaking Ratings scale in the context of primary school EFL learners in Indonesia. Content validity evidence was demonstrated through scale-level Aiken's V coefficients ranging from 0.667 (acceptable) to 1.000 (excellent), indicating good to very strong expert agreement regarding construct representation. Most dimensions, including descriptor clarity, test format appropriateness, and language suitability for children's contexts, demonstrated strong content validity (≥ 0.80). In terms of reliability, inter-rater reliability analysis using the two-way random effects ICC model with absolute agreement indicated moderate to good reliability, with ICC(2,1) values ranging from 0.703 to 0.835 and ICC(2,k) values ranging from 0.826 to 0.910. The relatively controlled SEM values (0.463–4.516) further support the interpretation that the scores demonstrate adequate measurement precision for classroom-based assessment contexts. These findings contribute to argument-based validity theory by demonstrating that quantitatively supported content representation, combined with empirically measured scoring consistency, can provide a foundation for evaluation inferences in children's speaking assessment. Practically, this study extends the existing literature, which has largely focused on adult populations, by providing empirical validity evidence in the context of Indonesian primary school EFL learners. However, these findings represent an initial stage of validation and are limited by the small-scale pilot design (N = 30), the involvement of only two raters, a single school context, limited monologue task sampling, and the absence of construct and criterion-related validity evidence.

Future research is recommended to employ Many-Facet Rasch Measurement to model student–rater–task interactions, Generalizability Theory to estimate multiple sources of measurement variance, and factor analysis or Item Response Theory (IRT) to strengthen construct validity evidence. Criterion-related evidence, such as correlations with other speaking proficiency measures, and measurement invariance testing across groups are also necessary to support broader generalizability. Practically, the adapted scale is recommended for formative assessment use with a minimum rater training package that includes construct briefing, exemplar video calibration, and moderation sessions, while its use for high-stakes summative assessment requires more comprehensive and advanced validation evidence.

REFERENCES

Acosta-Banda, A., Aguilar-Esteva, V., Patiño Ortiz, M., & Patiño Ortiz, J. (2021). Construction and Validity of an Instrument to Evaluate Renewable Energies and Energy Sustainability Perceptions for Social Consciousness. *Sustainability*, *13*(4), 2333. https://doi.org/10.3390/su13042333

Arafah, B., Room, F., Suryadi, R., B., L. O. M. I. H., Juniardi, Y., & Takwa. (2023). Character Education Values in Pullman's The Golden Compass. *Journal of Language Teaching and Research*, *15*(1), 246–254. https://doi.org/10.17507/jltr.1501.27

Arsyad, Moh. A., & Suadiyatno, T. (2024). Differentiated Assessment In EFL Classroom in Indonesia: Prospects and Challenges. *Journal of Language and Literature Studies*, *4*(2), 516–523. https://doi.org/10.36312/jolls.v4i2.1913

Chen, Z. (2025). What shapes communicative adequacy in second language speaking performance? The contributions of complexity, accuracy, fluency, and pronunciation. *Vigo International Journal of Applied Linguistics*, (22). https://doi.org/10.35869/vial.v0i22.4882

Đorđević, J. (2025). Rubrics in the Assessment of EAP Speaking Skills Supported by Mobile Assisted Language Learning. *ESP Today*, *13*(1), 91–112. https://doi.org/10.18485/esptoday.2025.13.1.5

Estrada Ramos, A. J., & Hernández Alipi, M. de los Á. (2025). Diferencias en la Experiencia de Aprendizaje del Inglés entre Niños y Adultos del CELE-UJAT. *Ciencia Latina Revista Científica Multidisciplinar*, *8*(6), 6227–6244. https://doi.org/10.37811/cl_rcm.v8i6.15318

Gao, J., & Sun, P. P. (2025). Unveiling the Relationship Between L2 Utterance Fluency and Perceived Fluency in Monologic and Dialogic Speaking. *Language and Speech*. https://doi.org/10.1177/00238309251352105

Gultom, C., Sihombing, R., & Harahap, S. H. (2024). Evaluasi Kemahiran Komunikasi Lisan Dalam Pembelajaran Bahasa Indonesia. *IJEDR: Indonesian Journal of Education and Development Research*, *2*(1), 445–448. https://doi.org/10.57235/ijedr.v2i1.1801

Halim, N., Kasim, N. A., & Pratiwi, D. F. (2025). DEVELOPING SPEAKING PROFICIENCY IN INDONESIAN EFL CLASSROOMS: A QUALITATIVE STUDY ON CHALLENGES AND SOLUTIONS. *THE ACADEMIC: ENGLISH LANGUAGE LEARNING JOURNAL*, *10*(1), 9–18. https://doi.org/10.52208/aellj.v10i1.1384

Huang, B. H., Bailey, A. L., Sass, D. A., & Shawn Chang, Y. (2021). An investigation of the validity of a speaking assessment for adolescent English language learners. *Language Testing*, *38*(3), 401–428. https://doi.org/10.1177/0265532220925731

Ikrima Maulida, Lestari, E., Kumala Sari, C., & Safutri, L. W. (2025). Formative Assessment as an Evaluation Tool for Elementary Students' Speaking Skills in Indonesian Language Learning: A Descriptive Qualitative Study. *Journal of Mathematics Instruction, Social Research and Opinion*, *4*(3), 769–782. https://doi.org/10.58421/misro.v4i3.610

Istihari, I., Juniardi, Y., Sofiah, V., & Abidin, Y. (2025). Text Complexity in An Indonesian EFL Textbook: Is it Aligned with the Emancipated Curriculum Goals? *Journal of English Language Studies*, *10*(1), 82. https://doi.org/10.30870/jels.v10i1.29104

Joo, D., & Lee, J. (2022). Validation of the L2 Speaking Performance Assessment for Young EFL Learners: Using Many-facet Rasch Measurement. *Korean Journal of Applied Linguistics*, *38*(3), 31–56. https://doi.org/10.17154/kjal.2022.9.38.3.31

Jung Youn, S. (2023). Test design and validity evidence of interactive speaking assessment in the era of emerging technologies. *Language Testing*, *40*(1), 54–60. https://doi.org/10.1177/02655322221126606

Kaharuddin, K., Arafah, B., Nurpahmi, S., Sukmawaty, S., Rahman, I. F., & Juniardi, Y. (2023). Exploring How Reading Aloud and Vocabulary Enrichment Shape English Speaking Skills Among Indonesian Learners of English. *World Journal of English Language*, *13*(8), 436. https://doi.org/10.5430/wjel.v13n8p436

Lane, S., & Marion, S. F. (2025). Validity Argumentation for Culturally Responsive Assessments 1. In *Culturally Responsive Assessment in Classrooms and Large-Scale Contexts* (pp. 106–123). Routledge. https://doi.org/10.4324/9781003392217-8

Lauwaert, P. (2023). On Validity. *Studies in Applied Linguistics and TESOL*, *23*(1). https://doi.org/10.52214/salt.v23i1.11804

Liao, M.-H. (2025). Cultivating proficient and efficacious L2 English speakers via VoiceThread-mediated self- and peer assessments. *Humanities and Social Sciences Communications*, *12*(1), 1277. https://doi.org/10.1057/s41599-025-05674-2

Mercado Borja, W. E., & Barrera Navarro, J. R. (2023). Diseño, construcción y validación de un instrumento que evalúa acciones innovadoras mediadas con TIC. *Sophia*, *19*(2). https://doi.org/10.18634/sophiaj.19v.2i.1287

Milano, N., Ponticorvo, M., & Marocco, D. (2026). Human Expertise and Large Language Model Embeddings in the Content Validity Assessment of Personality Tests. *Educational and Psychological Measurement*, *86*(1), 30–53. https://doi.org/10.1177/00131644251355485

Mister, B. (2025). Enhancing Adult ESL Learners' Vocabulary Use through Pronunciation-Focused Discussion. *Teaching English as a Second or Foreign Language--TESL-EJ*, *29*(2). https://doi.org/10.55593/ej.29114a4

Mulyanah, E. Y., Arwen, D., Ishak, Muhyidin, A., Nulhakim, L., Jamaludin, U., & Kumala, S. A. (2025). The Impact of E-Visual English Instructions Prototype, Local Wisdom, and Spiritual Values on Pre-Teachers' Perceptions and English Teaching Effectiveness. *Theory and Practice in Language Studies*, *15*(9), 3124–3135. https://doi.org/10.17507/tpls.1509.35

Mulyanah, E. Y., Juniardi, Y., & Nulhakim, L. (2025a). Revitalizing the Name Euis in Sundanese Culture Integrated E-Visual English Instructions (EUIS) in Implementing the Merdeka Curriculum. *3rd ISOLLEAC (International Seminar on Language, Literature, Educayion, Arts, and Culture) Available*, 1–10. https://doi.org/http://dx.doi.org/10.62870/aiselt.v10i2.37106

Mulyanah, E. Y., Juniardi, Y., & Nulhakim, L. (2025b). The Effectiveness of EUIS (E-Visual English Instructions) in Enhancing Primary School Teachers' English Skills. *PROCEEDING AISELT (Annual International Seminar on English Language Teaching)*, 385–392. https://doi.org/10.62870/aiselt.v10i1.36882

Neiriz, R. (2023). Developing and evaluating a contextualized interactional competence rating scale based on a metaphorical conceptualization. *Journal of Second Language Studies*, *6*(1), 61–94. https://doi.org/10.1075/jsls.22003.nei

Nguyen, C. D. (2021). The construction of age-appropriate pedagogies for young learners of English in primary schools. *The Language Learning Journal*, *49*(1), 13–26. https://doi.org/10.1080/09571736.2018.1451912

Nulhakim, L., Wibawa, B., & Erwin, T. N. (2019). Relationship between students' multiple intelligence-based instructional areas and assessment on academic achievements. *Journal of Physics: Conference Series*, *1188*(1). https://doi.org/10.1088/1742-6596/1188/1/012086

Ölmezer-Öztürk, E., & Aydin, B. (2018). Toward measuring language teachers' assessment knowledge: development and validation of Language Assessment Knowledge Scale (LAKS). *Language Testing in Asia*, *8*(1), 20. https://doi.org/10.1186/s40468-018-0075-2

Putri, A., & Sya, M. F. (2023). Tantangan Berbicara Bahasa Inggris pada Siswa Sekolah Dasar. *Karimah Tauhid*, *2*(2), 510–516. https://doi.org/10.30997/karimahtauhid.v2i2.7850

Quesada Pacheco, A. G. (2023). Assessment of Young English-Language Learners. *Revista de Lenguas Modernas*, (36). https://doi.org/10.15517/rlm.v0i36.48313

Raymond, J., Dai, D. W., & McAllister, S. (2025). The interpretation-use argument– the essential ingredient for high quality assessment design and validation. *Advances in Health Sciences Education*, *30*(4), 1313–1332. https://doi.org/10.1007/s10459-024-10392-6

Rima, R., Juniardi, Y., & Syafrizal, S. (2025). Assessing Self-Regulated Learning of Undergraduate EFL Students: Instrument Development and Validation. *International Journal of Social Learning (IJSL)*, *5*(2), 396–411. https://doi.org/10.47134/ijsl.v5i2.387

Schames Kreitchmann, R., Nájera, P., Sanz, S., & Sorrel, M. Á. (2024). Enhancing Content Validity Assessment With Item Response Theory Modeling. *Psicothema*, *36*(2), 145–153. https://doi.org/10.7334/psicothema2023.208

Shella Gherina Saptiany, & Bayu Ade Prabowo. (2024). Speaking Proficiency Among English Specific Purpose Students: A Literature Review On Assessment And Pedagogical Approaches. *LITERACY : International Scientific Journals of Social, Education, Humanities*, *3*(1), 36–48. https://doi.org/10.56910/literacy.v3i1.1392

Sinclair, J., & Lau, C. (2018). Initial assessment for K-12 English language support in six countries: revisiting the validity–reliability paradox. *Language and Education*, *32*(3), 257–285. https://doi.org/10.1080/09500782.2018.1430825

Speyer, R., Hakkarainen, A., Yoon, S., Kim, J.-H., Windsor, C., Wilkes Gillan, S., Littlefair, D., & Cordier, R. (2024). Content validity of measures in early numeracy in children up to eight years: A COSMIN systematic review. *PLOS ONE*, *19*(9), e0308874. https://doi.org/10.1371/journal.pone.0308874

Spoto, A. (2025). Supplemental Material for Improving Content Validity Evaluation of Assessment Instruments Through Formal Content Validity Analysis. *Psychological Methods*. https://doi.org/10.1037/met0000545.supp

Sultana Shaik, S. (2024). DEVELOPING COMMUNICATION PROFICIENCY: A MULTIDIMENSIONAL ANALYSIS OF LANGUAGE COMPETENCIES. *International Journal of Advanced Research*, *12*(05), 1144–1151. https://doi.org/10.21474/IJAR01/18829

Suzuki, S., & Kormos, J. (2025). L2 fluency across tasks: disentangling demands on conceptualisation and formulation in speech production. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2024-0185

Vo, S. (2017). Assessing Foreign Language Students' Spoken Proficiency: Stakeholder Perspectives on Assessment Innovation , By M. East. *Language Assessment Quarterly*, *14*(1), 93–96. https://doi.org/10.1080/15434303.2016.1262378

Widiastuti, O. (2025). Developing Assessment in Indonesian EFL Speaking Classroom. *Issues in Applied Linguistics & Language Teaching*, *7*(1), 305–317. https://doi.org/10.37253/iallteach.v7i1.10400

Zhang, X., & Lu, X. (2025). Aligning linguistic complexity with the difficulty of English texts for L2 learners based on CEFR levels. *Studies in Second Language Acquisition*, *47*(5), 1407–1434. https://doi.org/10.1017/S0272263125101125

Zhang, Y., & Zhang, L. J. (2024). Developing and validating an L2 writing willingness to communicate scale: A sequential embedded mixed-methods approach. *Language Teaching Research*. https://doi.org/10.1177/13621688241279834

Заирова, Н. (2023). Assessment methods for evaluating communication skills in english language learners. *Ренессанс в Парадигме Новаций Образования и Технологий в XXI Веке*, *1*(1), 460–464. https://doi.org/10.47689/XXIA-TTIPR-vol1-iss1-pp460-464