Indonesian Journal of Foreign Language Studies, Vol. 2 No. 1, 2025
e-ISSN: 3090-384X
https://indofes.org/index.php/journal-indofes

The Validity and Reliability Grade Primary School EFL Students' Final Speaking Test

Anis Zayyana
Universitas Negeri Surabaya
anis.23008@mhs.unesa.ac.id

Abstract

The study underscores the essential importance of mastering English speaking skills at an early age as mandated by Indonesia's Merdeka curriculum, which establishes English speaking as a vital subject for assessment in primary education. To ensure the effectiveness of the speaking test used for sixth-grade EFL students, the study focuses on evaluating the validity and reliability of the test. Using a qualitative methodology that combines interviews with an English teacher serving as the rater and document analysis, the research investigates multiple facets of test quality. The findings demonstrate that the speaking test is valid through the application of content validity—confirming the relevance and representativeness of test items—and consequential validity, which considers the implications and fairness of the test outcomes. Moreover, reliability is verified via intra-rater reliability, indicating that the scoring is consistent when conducted by the same evaluator over time. Ultimately, the study concludes that the speaking test for sixth-grade primary school EFL students possesses excellent quality in both validity and reliability. Consequently, it suggests that other educational institutions with similar levels and objectives may confidently adopt or adapt the test's design, scoring rubric, and rating procedures to improve their own assessment practices.

Keywords: *Validity, Reliability, Speaking, Test, EFL*

Submitted:
June 13th 2025

Revised:
June 29th 2025

Accepted:
June 30th 2025

How to cite (APA 7th)
Zayyana, A. (2025). The Validity and Reliability Grade Primary School EFL Students' Final Speaking Test. *Indonesian Journal of Foreign Language Studies, 2*(1), 19-27

1. Introduction

Among the four key language skills, speaking is considered as the most prominent skill to be mastered by language learners. Namaziandost et al., (2019) stated that a comprehensive understanding of English, particularly spoken English, is required in today's work environment. This is in line with the findings of Kojima & Fukui (2024), who revealed that university students are driven to learn English, particularly speaking, as a second or third language because it will provide them with more options in their future careers. Furthermore, many language learners place a high value on mastering speaking skills, and they frequently assess their success based on their advancement in speaking abilities (Riasati, 2018). To conclude, speaking is clearly the most crucial ability for language learners to develop.

For primary school EFL students, developing speaking skill is a critical objective in EFL learning. The study found that primary school L2 students aspired to speak and learn English because learning English would allow them to understand teachers more quickly (Gundarina, 2023). Moreover, Goriot & van Hout (2023) showed in their study that the impact of early English education on the development of communicative scope would be perceived more favorably by teachers from primary schools. In addition, oral language is the most important predictor of text comprehension in the first grade of school (Papadimitriou & Vlachos, 2014). It is obvious that speaking skill is prominent for primary students.

Speaking abilities for students in primary schools are also important to develop in Indonesia. English is regarded as a foreign language in Indonesia, yet it is a compulsory subject in formal education. In the current curriculum, known as Merdeka or Independent curriculum, the Indonesian Ministry of Education divided English learning into six phases (A-F) with various learning goals. Primary schools should complete three phases: A, B, and C (Badan Standar Kurikulum dan Asesmen Pendidikan, 2022). The first phase of language learning focuses on introducing and developing oral language skills. In phase B, however, English learning must remain focused on oral language while written language is gradually introduced. During phase C, the last phase in primary school, English learning must be focused both orally and in writing. Thus, it can be stated that, even in Indonesia, learning English from a young age is crucial, and the first ability to be cultivated is speaking, emphasizing the importance of speaking skills in primary school.

Since speaking skills are an essential part of the curriculum, this makes them an essential object of assessment as well. In the Merdeka curriculum there were two assessment concepts recommended (Badan Standar Kurikulum dan Asesmen Pendidikan, 2022b). The first concept is called formative assessment, and it aims to provide students with information and feedback to help them improve their learning process. Formative assessment is carried out at the beginning of a lesson to determine students' readiness to learn new information. Aside from that, it can also be carried out during the learning process to monitor students' progress and provide immediate feedback. The second notion is summative assessment, which aims to ensure that all learning objectives have been met. Summative assessment occurs at the end of the learning process. It is also possible to examine many learning objectives simultaneously. Summative assessment, unlike formative assessment, is incorporated into the assessment calculation at the end of each semester, academic year, or educational level.

According to Brown and Priyanvada (2019), one of the assessment methods in speaking to measure learners' ability, knowledge, or performance which occur at identifiable times in curriculum is named a test. To conduct a speaking test, the learners must be given a task to talk about (Luoma, 2004). Many experts have classified different types of speaking tasks. However, Brown and Priyanvada (2019) provide the simplest and most understandable classification. Brown and Priyanvada (2019) classified speaking tasks into four: imitative speaking task, intensive speaking task, responsive speaking task, interactive speaking task, and extensive speaking task. Imitative speaking tasks require the learner to repeat or imitate the speech. Whereas, intensive speaking tasks require the learner to produce language in controlled, specific settings, frequently with a focus on certain grammatical or lexical elements. Next, responsive speaking tasks require exchanges in which the learner responds to prompts or questions. Interactive speaking tasks involve lengthier, more complex conversations in which participants interact with one another, negotiating meaning and frequently solving issues or discussing topics. Last, extensive speaking tasks typically involve long monologues in which the learner speaks for an extended period of time, such as presenting presentations or recounting stories.

To aim for the usefulness of the speaking test' score, it is important to ensure the validity and reliability of the test. Luoma (2004) mentioned that speaking test results must, like other test results, be trustworthy, equitable, and most importantly, helpful for the intended uses. In this regard, testing experts primarily utilize two technical attributes: validity and reliability. Validity refers to the ratings' significance for the purposes for which they are designed, whereas reliability deals with the scores' consistency (Luoma, 2004).

Validity is by far the most difficult requirement for a test to be effective and possibly the most crucial one (Brown and Priyanvada, 2019). Luoma (2004) also stated the same that when developing tests, validity is the most crucial factor to take into account. According to Luoma (2004) there are several steps to ensure the validity of the speaking test. The first step is to define the purpose of the test. The next step is to choose which type of speaking test is meant to assess: linguistic, communicative, or task-based. Following that, demonstrate through the test development process that the tasks and criteria, as well as the administration and grading processes, effectively execute the construct. The evaluation of the rating criteria comes next. Lastly, every preparation and observation the test developers make about the usage of scores is included in the validation evidence. Whereas, based on Brown and Priyanvada (2019), there are five types of validity: content-related validity, criterion-related validity, construct validity, consequential validity, and face validity. Tests that sample the subject matter and require the test-taker to do the measured behavior can provide content-related validity. Criterion-related validity refers to the amount to which the test's criterion is actually achieved. Whereas, construct validity refers to whether a test accurately represents the theoretical construct as defined. Next, consequential validity emphasizes the potential significance of the outcomes of using a test. Last, face validity is the degree to which a test appears to measure the claimed knowledge or abilities, based on the subjective judgment of examinees, administrative personnel, and unsophisticated observers.

Reliability in the test means that the test is consistent and dependable. It implies that if the scores from a test given today are reliable, they will be nearly identical if the test is given to the same people again tomorrow (Luoma, 2004). Brown and Priyanvada (2019) also mentioned that the test should produce similar outcomes if it is administered to the same student or students who are matched on two separate times. Luoma (2024) suggested that the methods for ensuring reliability of formal tests and classroom tests are somewhat different. For formal tests such as the TOEFL and IELTS speaking test, the most common method to ensure reliability is by using rater training. Another method that is often used to ensure reliability is standard setting. For classroom tests, rater’s internal consistency is usually employed. Luoma (2024) also mentioned there are three types of reliability which are relevant for speaking test: intra-rater reliability, inter-rater reliability, and parallel form reliability. Whereas, according to Brown and Priyanvada (2019), there are four methods to ensure reliability: student-related reliability, rater reliability, test administration reliability, and test reliability.

Existing previous studies have already demonstrated the validity and reliability of speaking tests. A study conducted by Lu et al. (2016) aimed to investigate the validity and reliability of computer-assisted English speaking test. The participants were 34 non-English major university students. An experiment was carried out in a digital language lab to determine the effectiveness of the speaking exam format and its educational implications. The participants took part in an English Audio-video Speaking Course (EAVSC) held in a digital language lab. SPSS software was used to conduct statistical analysis to determine the effectiveness of the test. The results showed that the computer-assisted English speaking test is valid and reliable, with a favorable impact on teaching and learning in the EAVSC environment.

Another study by Huang et al. (2020) examined the construct validity of the TOEFL Junior speaking test for adolescent EFL learners in Taiwan. After the quantitative analysis conducted, the findings provide significant evidence for the test's internal structure and positive relationships between test scores and external factors, which support the test's construct validity. Other than that, Xu et al., (2021) have investigated the reliability of an automated scoring of learner speech in an online oral English test. Despite having great internal consistency, the study discovered that the automarker was marginally lenient than examiner fair average ratings, especially for low-proficiency speakers. Additionally, Koizumi (2022) also has investigated the speaking assessment in secondary classrooms in Japan. It resulted that although L2 speaking assessments in secondary school classes in Japan should be undertaken on a regular and adequate basis, and used summative and formative, they are not widely practiced. There are numerous issues concerning teacher-made, teacher-scored speaking assessment. The findings concluded that the validity and reliability of speaking assessment is not ensured. Koizumi (2022) also provided several future directions to enhance the quality of speaking assessment. Khan et al. (2022) also have assessed the inter-rater reliability of a speaking test of Saudi EFL undergraduate learners during remote learning. Khan et al. (2022) used statistical methods such as correlation coefficients and the Bland-Altman test to assess raters' agreement. The Bland-Altman test showed that the speaking test is dependable and can provide a good evaluation of speaking skill.

Previous studies on the validity and reliability of speaking tests have focused on adolescents or university students. There is little research on the validity and reliability of speaking tests in primary school kids, particularly in sixth grade. Besides, the majority of studies on the validity and reliability of speaking tests utilizes quantitative methods. More qualitative study is needed to gain a deeper understanding of teachers' points of view. Therefore, the research question for this study is: How valid and reliable is the final speaking test for sixth-grade EFL students?

2. Method

This study employed two techniques to gather the data, interview and document analysis. The first technique used was an interview. In-depth interview was held with an English teacher from a private primary school in Pasuruan, East Java, Indonesia. The English teacher served as an examiner and rater, assessing and scoring the students' final speaking test. In the first interview session, the English teacher was asked to recount how the final English-speaking test for 6th grade primary school EFL students was conducted. This was followed by asking further questions about her experience as a rater in rating the speaking test. In the next interview session, the English teacher was asked to discuss the purpose of the speaking test and how to maintain consistency in rating students’ speaking test. The two sessions were recorded, and the recordings were transcribed and translated for analysis. The translations were made to be as close to the original text as possible. The second technique used for gathering data was document analysis. Several documents to be analyzed were curriculum documents, lesson plans, and scoring rubric. The curriculum documents are the documents given by the Ministry of Education consisting of learning objectives of every phase (A-F), and guidance for conducting assessment. Lesson plans and scoring rubric was given by the English teacher as a participant of this study.

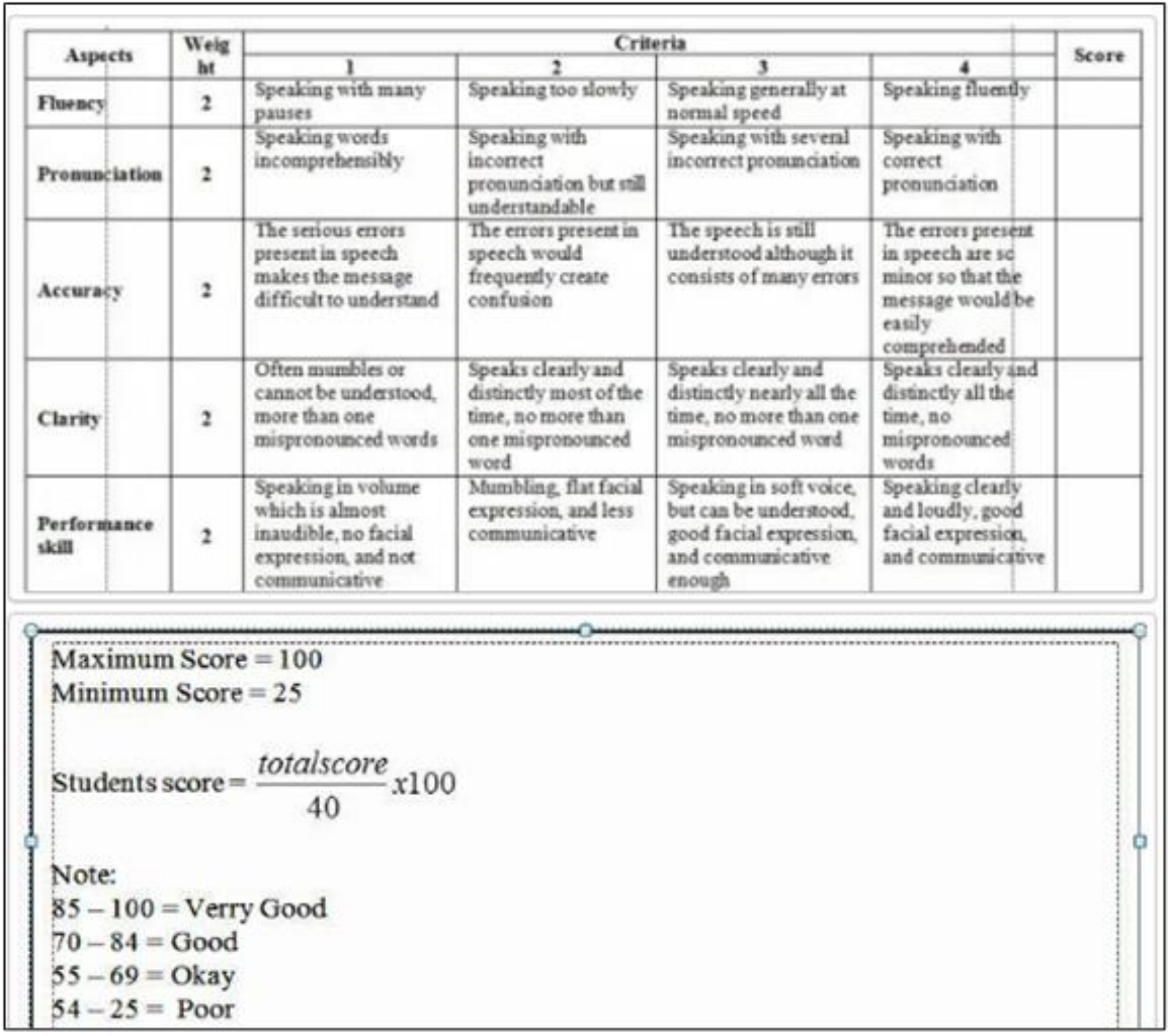
The data about validity of the speaking test were analyzed and interpreted using Brown H. Douglas & Abeywickrama Priyanvada (2019)’s content validity and consequential validity checklist. Whereas, the data about reliability of the speaking test were analyzed using Luoma (2004) and Brown & Priyanvada (2019)’s notion.

3. Findings and Discussions

The Final Speaking Test of 6th Grade Primary School EFL Students

The speaking test aimed to measure 6th grade EFL students' ability in speaking. The test was constructed by the association of English teachers in Pasuruan, East Java, Indonesia. The speaking test was intended to measure all student’s ability, knowledge, and performance in a phase C for 6th grade primary which is stated in Indonesia’s Merdeka curriculum. The 6th grade students were asked to introduce themselves in the final speaking test.

The scoring rubric for the speaking test was adopted from Susanti et al. on their personal blog which the source of it is still unknown. However, the scoring rubric then was analyzed and discussed again by the association of English teachers in Pasuruan, East Java, Indonesia. The following are the scoring rubric used for rating speaking test.



Aspects	Weig ht	Criteria				Score
		1	2	3	4	
Fluency	2	Speaking with many pauses	Speaking too slowly	Speaking generally at normal speed	Speaking fluently	
Pronunciation	2	Speaking words incomprehensibly	Speaking with incorrect pronunciation but still understandable	Speaking with several incorrect pronunciation	Speaking with correct pronunciation	
Accuracy	2	The serious errors present in speech makes the message difficult to understand	The errors present in speech would frequently create confusion	The speech is still understood although it consists of many errors	The errors present in speech are so minor so that the message would be easily comprehended	
Clarity	2	Often mumbles or cannot be understood, more than one mispronounced words	Speaks clearly and distinctly most of the time, no more than one mispronounced word	Speaks clearly and distinctly nearly all the time, no more than one mispronounced word	Speaks clearly and distinctly all the time, no mispronounced words	
Performance skill	2	Speaking in volume which is almost inaudible, no facial expression, and not communicative	Mumbling, flat facial expression, and less communicative	Speaking in soft voice, but can be understood, good facial expression, and communicative enough	Speaking clearly and loudly, good facial expression, and communicative	

Maximum Score = 100
Minimum Score = 25

$$\text{Students score} = \frac{totalscore}{40} x100$$

Note:
85 – 100 = Verry Good
70 – 84 = Good
55 – 69 = Okay
54 – 25 = Poor

Figure 1. The Scoring Rubric

The final speaking test took place in one of Pasuruan’ s private primary schools in East Java, Indonesia, with 18 students. Their English teacher served as both examiner and rater for the test. The test was performed in their classroom. The students entered the classroom one by one, while the rest waited outside. All of the students took the final speaking test on the same day. The students' turns in the final speaking test were decided randomly using an online spinning wheel application rather than their attendance book. All of the students took part in the decision-making process and agreed on it.

The Validity of Speaking Test

The rater explained that the speaking test was conducted since it was one of the mandatory requirements that sixth-grade primary school students must complete in order to pass and graduate. Furthermore, the decision to use self-introduction as a topic for a speaking test was made during a forum discussion among English teachers in Pasuruan, East Java, Indonesia. It was chosen since the topic was taught in sixth grade and was mentioned in the curriculum document in phase C. Also, it will be addressed again in seventh-grade junior high school. Therefore, it was a wise option to choose this topic as a bridge from primary to junior high school. The rater said:

Choosing self-introduction as a topic for a speaking test is an excellent choice. Self-introduction is one of the topics covered in primary school. It is beneficial for students to revisit *their previous knowledge, what they have learned. Besides, it was meant to prepare students. I mean, in junior high school, students will be asked to introduce themselves in English.*

The table below showed the detailed result of validity analysis of the 6th grade primary EFL students' final speaking test.

Table 1. The Results of Validity Analysis

Aspects of Validity	***Findings***	***Detail Descriptions***
Content Validity	Phase C' objectives (listening-speaking) in Merdeka curriculum clearly identified in the speaking test	*"............. Students use simple English to interact and communicate in familiar/usual/routine situations. ………"* *"By the end of Phase C, students use English to interact in a range of predictable social and classroom situations using certain patterns of sentences. …………"* *"...................Students understand the relationship between letter sounds in simple vocabulary in English………."*
	Test specifications have embedded in assessment documents or guidelines	*"The test specifications consist of the content coverage, time and administration, scoring rubric, etc. appeared in assessment guidelines provided"*
	Test specifications include task that have already been performed as part of the course procedures	*"Students have already taught and practiced how to introduce the mselves."*
	Test specifications include task that represent most of the objectives of phase C in Merdeka curriculum	*"It stated that the topic for the final speaking test is self-introduction which actually unintentionally has already covered most of the objectives of phase C in Merdeka curriculum."*
	The task involved actual performance of target task(s)	*"Self-introduction topic in the speaking test really reflected well on the activity that the students might encounter in real life context."*
Consequential validity	Offered students appropriate review and preparation for the test Suggested test-taking strategies that will be beneficial	*"Before the speaking test was conducted, I, as the test takers' English teacher, had already reviewed all the material taught. Also, before they entered the class to perform the test, I gave them a clear and thorough explanation about the test."*
	The weaker students were not be overwhelmed, but the best students were slightly challenged	*"This test might seem simple and easy for several students, but actually not that difficult for the low achievers."*
	The test provided beneficial washback	*"For me, after the students' speaking test, I adapt and change a particular way of how I teach students. I also give students feedback that can help them develop."*
	Students encouraged to see the test as a learning experience	*"My students said that they learned and experienced new things during the final speaking test."*

The Reliability of Speaking Test

Prior to the test, the rater ensured that there was no noise in the classroom or surrounding locations. Following that, the rater clearly explained to all of the students how the test will be administered on that day. The rater also presented the students with the scoring rubrics, describing what aspects would be scored and what they should do to attain a specific score. In addition, before administering the test, the rater ensured that all of the students understood the entire procedure.

When the test was conducted, the rater rated the students directly based on the criteria outlined in the scoring rubric. To determine appropriate scores for the students, the rater followed guidelines established by the Pasuruan English Teachers Association in East Java, Indonesia.

Other than that, the test was also recorded using two cameras. According to the rater, in addition to rating students directly while they took the test, the test was also videotaped using two cameras, one recording students from the front and the other recording students and the rater from the back. The rater mentioned:

I am convinced that recording the process of the speaking test will help me maintain the consistency of the scores when rating the students.

The rater further stated that the students' scores were verified after double-checking the video footage of the test. The double-checking session resulted in two students’ scores being modified. The rater revealed:

To prevent inconsistencies in student scores, I watched the recordings of the students when they took the final speaking test. Then I discovered that two students' scores were not suitable for their performance. So I decided to make changes to the scores after reviewing the videos again.

Discussion

The final speaking test conducted by 6th grade students in one of the private primary schools in Pasuruan, East Java, Indonesia has already followed the assessment guideline provided by the Ministry of Education in Indonesia. It was noticeable that the speaking test was part of summative assessment to measure whether the learning objectives stated in phase C Merdeka curriculum have already been achieved or not at the end of the learning process. Other than that, the final speaking test’ results were also used as a part of assessment for passing the educational level (primary). That also aligns with the Ministry of Education in Indonesia suggested about the summative assessment.

To ensure the validity of the speaking test, the content validity and consequential validity was employed. This is relevant with Brown H. Douglas & Abeywickrama Priyanvada (2019) notion about applying principles in classroom testing. From the findings, it is known that the decision to choose self-introduction as a topic for the speaking test derived from the learning objectives of phase C in Merdeka curriculum. Also, inside the self-introduction, the students practiced their ability, knowledge, and performance in a phase C for 6th grade primary which is stated in Indonesia’s Merdeka curriculum. This is in line with what Luoma (2004) said that the very first step to ensure validity is by clearly clarifying the purpose of the test. Other than that, the scoring rubric employed was an analytic rubric that encompassed a variety of aspects, the majority of which were linguistic. Thus, the speaking construct is the linguistic approach. It shows that the second step of ensuring the validity of the speaking test based on Luoma (2004) which is defining the test construct has already been done. Besides, The task (self-introduction) engages examinees in similar spoken interaction as in non-test situations. It shows evidence that the test implements the construct which, as Luoma (2004) said. Furthermore, the aspects stated in the scoring rubric aligned with the learning objectives of phase C in Merdeka curriculum. This pinpoints that it has already followed the fourth step of Luoma (2004). The test also resulted in beneficial washback which also aligns with the last step of ensuring validity by Luoma (2004). To sum up, the final speaking test for 6th grade primary school EFL students is valid.

To determine the reliability of the speaking test, the intra-rater reliability was used. The use of this type of reliability is relevant to what Luoma (2004) stated that the basic way to ensure the consistency of ratings for classroom assessments is by using rater's internal consistency. To avoid the subjectivity, the rater rated speaking test performances immediately or one task at a time. This is again in line with one of Luoma (2004) and Brown H. Douglas & Abeywickrama Priyanvada (2019) suggestions for reducing subjectivity which is often raised in classroom assessments. The rater also reflected the rating work by revisiting the students' performances that have videotaped after finishing the rating of the last performance. The rater' activity in accordance with what Luoma (2004) mentioned about simple self-check of consistency that can be done in classroom assessments. The rater also used analytic scoring rubric to rate students in the final speaking test. It is mentioned in Brown H. Douglas & Abeywickrama Priyanvada (2019) that one of the ways to increase intra-rater reliability is by applying careful specification of an analytical scoring rubric. Thus, it can be concluded that the final speaking test for 6th grade primary school EFL students is reliable.

4. Conclusion and suggestion

The results of this study demonstrated that the final speaking test for 6th-grade primary school EFL students is both valid and reliable, indicating that the test possesses high quality and effectiveness. Moreover, the test's procedures, scoring rubric, and rating methods are well-designed and can be adopted or adapted by other educational institutions aiming to assess speaking skills at the same educational level and for similar purposes. It is important to note that this study focused exclusively on evaluating the validity and reliability of the final speaking test applied to 6th-grade primary students following Phase C of the Merdeka Curriculum in Indonesia. Future research should explore other aspects of assessment principles, different curriculum phases, and various educational levels to provide a more comprehensive understanding of speaking test evaluation and applicability.

References

Brown H. Douglas & Abeywickrama Priyanvada. (2019). *Language Assessment: Principles and Classroom Practices* (Third edit). Pearson Education.

Goriot, C., & van Hout, R. (2023). Primary-school teachers' beliefs about the effects of early-English education in the Dutch context: communicative scope, disadvantaged learning, and their skills in teaching English. *International Journal of Bilingual Education and Bilingualism*, *26*(4), 498–513. https://doi.org/10.1080/13670050.2022.2124841

Gundarina, O. (2023). Migrant pupils as motivated agents: a qualitative longitudinal multiple-case study of Russian-speaking pupils' future ideal selves in English primary schools. *The Language Learning Journal*, *51*(6), 749–765. https://doi.org/10.1080/09571736.2022.2073383

Huang, B. H., Bailey, A. L., Sass, D. A., & Shawn Chang, Y. (2020). 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

Khan, R. M. I., Kumar, T., Benyo, A., Jahara, S. F., & Haidari, M. M. F. (2022). The Reliability Analysis of Speaking Test in Computer-Assisted Language Learning (CALL) Environment. *Education Research International*, *2022*(1), 8984330. https://doi.org/https://doi.org/10.1155/2022/8984330

Koizumi, R. (2022). L2 Speaking Assessment in Secondary School Classrooms in Japan. *Language Assessment Quarterly*, *19*(2), 142–161. https://doi.org/10.1080/15434303.2021.2023542

Kojima, N., & Fukui, H. (2024). L2 English and L3 Japanese motivation, international posture, and success of students in an English-medium instruction (EMI) program at a Japanese University. *Journal of Multilingual and Multicultural Development*, 1–17. https://doi.org/10.1080/01434632.2024.2342925

Lu, Z., Li, Z., & Hou, L. (2016). *On the Validity and Reliability of a Computer-assisted English Speaking Test BT - Proceedings of the 2016 International Conference on Intelligent Control and Computer Application*. 187–193. https://doi.org/10.2991/icca-16.2016.43

Luoma, S. (2004). *Assessing Speaking*. Cambridge University Press.

Namaziandost, E., Neisi, L., Kheryadi, & Nasri, M. (2019). Enhancing oral proficiency through cooperative learning among intermediate EFL learners: English learning motivation in focus. *Cogent Education*, *6*(1), 1683933. https://doi.org/10.1080/2331186X.2019.1683933

Papadimitriou, A. M., & Vlachos, F. M. (2014). Which specific skills developing during preschool years predict the reading performance in the first and second grade of primary school? *Early Child Development and Care*, *184*(11), 1706–1722. https://doi.org/10.1080/03004430.2013.875542

Riasati, M. J. (2018). Willingness to speak English among foreign language learners: A causal model. *Cogent Education*, *5*(1), 1455332. https://doi.org/10.1080/2331186X.2018.145533

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