Jurnal Pendidikan Bahasa
Volume 15. Nomor 3, September 2025 | ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

Analysis of English Skill Tests for Grade VII Students Based on Language Assessment Principles

Yolanda Jesica Sitanggang[1], Tessaria Simorangkir[1], Angel Lamtama Sihaloho[1], Rotua Sinta Marito Pakpahan[1], Dumaris E Silalahi[1,*]

[1]Universitas HKBP Nommensen Pematangsiantar

*Corespondence: dumaris.silalahi@uhnp.ac.id

Abstract

This study evaluates English assessment instruments for Grade VII students at SMP Swasta Taman Asuhan based on five principles of language assessment: practicality, reliability, validity, authenticity, and washback. A mixed-method research design was employed, combining quantitative analysis of test validity and reliability with qualitative evaluation through classroom observations. The study involved 20 students and assessment tasks for writing, reading, listening, and speaking. The findings revealed that practicality was achieved as all tests were implemented efficiently using available resources and time. Reliability analysis showed high internal consistency for speaking and listening (Cronbach's Alpha = 0.94), writing (0.86), and moderate reliability for reading (0.59). Validity analysis using Pearson Product Moment indicated that most essay items in all skills were valid, while several multiple-choice items in writing and reading were invalid. Authenticity was evident through tasks relevant to students' real-life experiences. Washback was positive, as the assessments encouraged meaningful language use and provided valuable instructional feedback. These results affirm that the instruments are generally effective but require revision in certain areas.

Keywords: Language Assessment; Validity; Reliability; Authenticity; English Skill

Received: 26 Jun 2025; **Revised:** 13 Agu 2025; **Accepted:** 13 Agu 2025; **Available Online:** 16 Agu 2025

INTRODUCTION

Assessment plays a crucial role in English language teaching, not only to measure student achievement but also to improve instruction. Beyond assigning scores, it helps teachers make informed decisions, supports student development, and ensures fair evaluation of learning outcomes. For junior high school students in the early stages of second language acquisition, accurate and effective assessment tools are essential to measure progress constructively. A good language assessment should meet five essential principles: practicality, reliability, validity, authenticity, and washback (Brown, 2004). These principles ensure that tests are feasible, consistent, meaningful, relevant to real-life contexts, and capable of fostering positive educational impacts.

This study was conducted at SMP Swasta Taman Asuhan, Pematangsiantar, involving 20 students from Grade VII-1. The primary aim was to evaluate the English assessment instruments used for the four major language skills—writing, reading, listening, and speaking—to determine whether they align with the five principles of assessment. The assessment items included both multiple-choice and essay formats. To evaluate the validity of each item, the researcher used the Pearson Product–Moment correlation formula, as suggested by (Arikunto, 2012) cited in (Masrur & Kuning, 2025), to analyze the relationship between individual item scores and total test scores. This statistical method calculates the strength of the association between two variables, in this case, each test item and the overall score, to determine whether the items measure the intended skill. Meanwhile, reliability was measured using Cronbach's Alpha, a method recommended by (Sugiyono, 2017) and (Gay et al., 2012) cited in (Wibowo et al, 2021), that assesses the internal consistency of test items—meaning how closely related the items are as a group. A high Cronbach's Alpha value indicates that the items consistently measure the same construct.

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

In addition to quantitative methods, the study also involved qualitative analysis to assess the test's practicality, authenticity, and washback. Practicality was evaluated based on time management, resources, scoring ease, and student accessibility (Brown, 2004). Authenticity was examined by analyzing whether test tasks resembled real-world language use (Bachman & Palmer, 1996) cited in (Pusey, 2024), and washback was assessed based on the test's impact on teaching practices and student motivation (Alderson & Wall, 1993) cited in (Alqahtani, 2021). The findings showed that most of the test items, especially in listening and speaking, had high validity and very high reliability, with Cronbach's Alpha values reaching 0.94. While reading had the lowest reliability score (0.59), other skills such as writing and speaking showed strong psychometric quality. Furthermore, the assessments were found to be practical and authentic, and they also generated positive washback, encouraging students to engage more actively in learning.

The research was guided by the following problem: To what extent do the assessments for writing, reading, listening, and speaking skills meet the five principles of language assessment? In line with this, the objective of the research was to evaluate the quality of the English assessment instruments used for Grade VII students in terms of practicality, reliability, validity, authenticity, and washback.

This research is expected to provide both theoretical and practical contributions. Theoretically, it aims to enrich the literature on language assessment by providing empirical findings on how the five principles are applied in secondary education. Practically, the results may benefit English teachers by offering insights into the quality of their assessment tools; help students by ensuring fair and meaningful testing; and support schools and curriculum developers in improving their language evaluation frameworks.

Effective language assessment is guided by five interrelated principles: practicality, reliability, validity, authenticity, and washback. These principles, adapted from Brown and Abeywickrama (2019) cited in (Suryanto, 2025) and expanded by more recent scholars, serve as a framework for evaluating the quality and impact of language tests in real classroom settings.

Practicality refers to the feasibility of administering, scoring, and managing assessments within the constraints of school environments. Handayani and Susanti (2022) stress that assessments must be time-efficient, cost-effective, and manageable for teachers—especially in Indonesian schools where large class sizes and limited infrastructure are common. Practical formats like structured group speaking tasks or multiple-choice items are often preferred, provided they still capture relevant learning outcomes.

Reliability, as explained by Saputra and Fitriani (2023), emphasizes consistency in test results across different conditions, raters, and administrations. In productive skill assessment—such as speaking or writing—achieving reliability requires the use of clear rubrics and rater training. The goal is to minimize subjectivity and ensure fairness, so that a student's performance is judged consistently regardless of who assesses it.

Validity is the extent to which a test truly measures what it intends to measure. Wijaya and Salim (2021) highlight the importance of content, construct, and face validity in aligning test items with curriculum goals and learners' actual abilities. For example, a reading comprehension test must go beyond grammatical recognition and assess skills like inference and vocabulary interpretation to ensure its validity.

Authenticity, discussed by Astuti and Prasetyo (2024), refers to the degree to which assessment tasks reflect real-world communication. Tasks such as writing a social media post, giving directions, or responding to everyday situations enhance learners' engagement and motivation. Authentic assessments connect learning with meaningful language use, promoting deeper understanding and better performance.

Washback describes the influence of assessment on teaching practices and learner behavior. Pratiwi and Nugroho (2023) assert that assessments that encourage creativity, communication, and interaction foster positive washback. Conversely, overly rigid or test-driven formats may limit student growth. Thus, thoughtful test design can encourage both effective teaching strategies and student autonomy.

Taken together, these five principles must be balanced to create assessments that are pedagogically sound, ethically responsible, and practically applicable. A test that excels in one area but fails in another may not be effective overall. In this study, the assessment tools used at SMP Taman Asuhan—including multiple-choice and essay formats—are critically evaluated through this five-principal lens to determine their strengths and areas for improvement.

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

Literature Review

Concepts and Principles of Language Assessment

Language assessment in recent years has evolved into a dynamic and reflective component of language education. It is no longer viewed solely as a tool for measuring student achievement, but as a formative process that guides instruction, promotes student growth, and supports curriculum alignment.

According to Winna and Sabarun (2023), language assessment involves both standardized testing and classroom-based observations. This dual approach emphasizes the integration of formal and informal methods to obtain a more authentic picture of learners' language competence. Their view aligns with the growing emphasis on meaningful, communicative assessment practices in modern classrooms.

Language Assessment Literacy (LAL)

Building on this, Giraldo (2022) introduces the concept of Language Assessment Literacy (LAL), which stresses that teachers must possess theoretical, technical, and ethical knowledge to design effective assessments. Without sufficient LAL, even well-intentioned assessments may produce misleading or harmful outcomes. LAL equips educators to create fair, valid, and student-centered evaluations that reflect actual language use.

Lee and Butler (2020), as cited in Abrar-ul-Hassan and Nassaji (2024), argue that assessment should be understood as a dynamic system involving theory, practice, and professional reflection. They view LAL as essential in diverse educational contexts like Indonesia, where teachers must adapt assessments to meet varied student needs while staying aligned with educational reforms.

The Role of Technology in Language Assessment

The role of technology in language assessment has become increasingly significant, particularly after the COVID-19 pandemic. Jabbar (2023) discusses how tools such as AI-driven scoring systems, digital portfolios, and speech recognition can improve the practicality and validity of assessments. These innovations facilitate real-time feedback and support authentic assessment tasks, making testing more relevant to real-world communication.

Language Assessment in the Indonesian Educational Context

Locally, Astuti and Prasetyo (2024) define assessment as the purposeful design of tasks, while evaluation refers to the interpretation and application of results. Their work emphasizes the value of authentic assessments, such as writing emails or delivering short speeches, which mirror real-life communication and support communicative goals.

In the Indonesian setting, Wijaya and Salim (2021) underline the importance of balancing formative and summative assessments to meet national curriculum standards. Continuous feedback and final evaluations are both necessary to enhance instructional quality and student learning. Furthermore, Pratiwi and Nugroho (2023) highlight those well-designed assessments can generate positive washback, motivating students to learn meaningfully and encouraging teachers to focus on communication skills rather than rote memorization.

Assessment of Language Skills

In English language education, the four foundational skills—listening, speaking, reading, and writing—are central to communicative competence. These are typically classified into receptive skills (listening and reading) and productive skills (speaking and writing), each requiring distinct assessment approaches: a) Listening is a cognitively active process that combines bottom-up skills (recognizing sounds and syntax) with top-down strategies (prediction and inference). Kang and Wu (2021) recommend diverse tasks such as sequencing and open-ended responses to assess comprehension effectively, with authenticity and practicality being key considerations; b) Speaking is dynamic and interactional, involving fluency, accuracy, vocabulary, and appropriateness in real-time communication. Hartati and Siahaan (2023) stress that speaking assessments, including role plays and presentations, should be supported by analytic rubrics, standardized criteria, and trained raters to ensure reliability and fairness; c) Reading requires both literal and inferential comprehension. Marzuki and Rukmini (2022) emphasize assessing higher-order skills such as critical thinking, in addition to surface understanding. Reading tests may use multiple-choice and cloze tasks, and should also address digital

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

literacy for navigating online texts and media; d) Writing is an integrative skill that involves grammar, organization, coherence, and the expression of ideas. Lestari and Hamdani (2020) recommend using clear rubrics and realistic writing tasks—such as composing messages or reflections—that reflect authentic purposes. Formative feedback plays a critical role in helping students improve their writing over time.

These four skills, while interconnected, pose unique assessment challenges. Receptive skills tend to be easier to score objectively but may not capture depth of understanding, while productive skills yield richer data but require greater effort and judgment to assess. Consequently, balanced assessment approaches that integrate objective and subjective methods, as well as formative and summative evaluations, are essential. In the context of Indonesia's Kurikulum 2013, assessment emphasizes integrating these skills within communicative and character-building tasks. This study examines the instruments used at SMP Taman Asuhan to evaluate how effectively they measure each skill in alignment with the five principles of assessment: practicality, reliability, validity, authenticity, and washback.

METHOD

This study employed a mixed-method research design, combining both quantitative and qualitative approaches to obtain a comprehensive analysis of English assessment instruments based on the five principles of language assessment: practicality, reliability, validity, authenticity, and washback. The quantitative approach was used to evaluate the validity and reliability of the multiple-choice and essay items through statistical calculations, following procedures recommended by Arikunto (2012) and Sugiyono (2017), while the qualitative approach was used to analyze practicality, authenticity, and washback through direct observation and descriptive interpretation (Creswell & Plano Clark, 2018) cited in (Riazi & farsani, 2024). This research was conducted as a descriptive case study, focusing on the real application of assessment instruments in a specific classroom setting to capture in-depth and contextual insights.

The research took place at SMP Swasta Taman Asuhan, located in Pematangsiantar, involving a total of 20 students from class VII-1. The participants were selected through purposive sampling, considering their availability and willingness to participate after completing their final semester examinations. Among the participants, there were 12 male students and 8 female students, all of whom were in their early stage of English language acquisition and represented a typical population of junior high school learners in Indonesia.

The instruments used in this study included four sets of English language tests designed to assess students' competencies in writing, reading, listening, and speaking. Each test consisted of multiple-choice (MC) questions and essay-type items, allowing for both objective and subjective evaluation of student performance. The multiple-choice items focused on discrete aspects such as grammar, vocabulary, and comprehension, while the essay items evaluated students' ability to construct meaning through written or spoken production. To ensure fairness and consistency, analytic scoring rubrics were applied in evaluating essay and speaking tasks, with criteria such as fluency, organization, grammar, vocabulary use, and coherence. Additionally, a simple observation sheet was used to document the students' level of engagement and response during the testing process.

The data collection was conducted over a series of four separate class sessions, each allocated to assess one of the four language skills. The tests were administered after the students had completed their final semester exams, not during regular classroom instruction, in order to avoid disruption to the academic schedule. During the data collection period, an observation was conducted for one day, aiming to document how students interacted with the tasks, managed the instructions, and demonstrated engagement. No interviews or additional qualitative instruments were employed beyond this classroom observation.

The data obtained were analyzed using both quantitative and qualitative methods. For the quantitative analysis, Pearson Product Moment correlation was used to measure the validity of each test item, and Cronbach's Alpha was applied to determine the reliability of each section. These calculations were performed using Microsoft Excel. Descriptive statistics such as mean scores and percentages were also used to interpret student performance across different skills. For the qualitative analysis, a thematic interpretation was conducted based on the observational data, particularly in evaluating the aspects of practicality, authenticity, and washback as reflected through students' test behaviors and responses.

The research was carried out with the full approval and support of the school authorities, including the

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

principal and English teacher. Students participated voluntarily and were informed about the purpose of the study prior to the assessment. No pressure was applied, and all student responses were kept confidential. The data collected were used solely for academic purposes, ensuring adherence to ethical standards in educational research.

RESULT AND DISCUSSION

Data Analysis

To determine the overall test score, the weighted average of the multiple-choice and essay sections was computed. In this study, the multiple-choice section was weighted at 60% and the essay section at 40%, as essay responses required more subjective evaluation. The final score was calculated using this formula:

$$\text{Total Score} = (\text{MC Score} \times 0.6) + (\text{Essay Score} \times 0.4)$$

Using the previous example:

$$(83.33 \times 0.6) + (93.33 \times 0.4) = 87.33$$

This method ensured a balanced evaluation between objective and subjective measures and provided a fair reflection of students' language proficiency.

Table 1. Student Performance in English Skill Tests by Section (Writing, Reading, Listening, Speaking)

Name	Writing (MC / Essay / Total)	Reading (MC / Essay / Total)	Listening (MC / Essay / Total)	Speaking (MC / Essay / Total)
ASS	70.00 / 82.67/ 75.07	86.67 / 85.33 / 86.13	96.67 / 86.67 / 92.07	96.67 / 86.67 / 92.07
SBAP	73.33 / 80.00 / 75.60	76.67 / 84.00 / 80.33	90.00 / 86.67 / 88.67	90.00 / 86.67 / 88.67
AARH	66.67 / 97.33 / 78.13	66.67 / 94.67 / 77.33	80.00 / 98.67 / 87.47	80.00 / 98.67 / 87.47
Z	76.67 / 85.33 / 80.13	76.67 / 80.00 / 77.33	100.00 / 98.67 / 99.47	100.00 / 98.67 / 99.47
YM	66.67 / 88.00 / 74.80	86.67 / 80.00 / 84.00	60.00 / 98.67 / 75.47	60.00 / 98.67 / 75.47
MSN	66.67 / 92.00 / 76.40	80.00 / 88.89 / 83.56	90.00 / 100.00 / 94.00	90.00 / 100.00 / 94.00
WCS	70.00 / 80.00 / 74.00	66.67 / 80.00 / 72.00	83.33 / 100.00 / 90.00	83.33 / 100.00 / 90.00
YEK	46.67 / 80.00 / 60.00	100.00 / 80.00 / 92.00	96.67 / 100.00 / 98.67	96.67 / 100.00 / 98.67
FDR	73.33 / 92.00 / 80.27	80.00 / 85.33 / 82.13	86.67 / 100.00 / 92.00	86.67 / 100.00 / 92.00
RV	66.67 / 80.00 / 72.00	96.67 / 80.00 / 90.00	100.00 / 98.67 / 99.47	100.00 / 98.67 / 99.47
RF	46.67 / 80.00 / 60.00	56.67 / 71.11 / 62.84	60.00 / 100.00 / 76.00	60.00 / 100.00 / 76.00
MFS	20.00 / 17.78 / 18.67	56.67 / 71.11 / 62.84	46.67 / 98.67 / 67.47	46.67 / 98.67 / 67.47
IDS	40.00 / 97.33 / 62.93	83.33 / 71.11 / 78.20	40.00 / 44.44 / 41.78	40.00 / 44.44 / 41.78
ST	50.00 / 97.33 / 68.93	66.67 / 88.89 / 74.16	66.67 / 85.33 / 73.13	66.67 / 85.33 / 73.13
DNS	50.00 / 85.33 / 64.13	53.33 / 85.33 / 66.13	70.00 / 98.67 / 81.47	70.00 / 98.67 / 81.47
MFM	46.67 / 88.89 / 63.16	93.33 / 100.00 / 96.00	73.33 / 100.00 /	73.33 / 100.00 /

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

Name	Writing (MC / Essay / Total)	Reading (MC / Essay / Total)	Listening (MC / Essay / Total)	Speaking (MC / Essay / Total)
			84.00	84.00
MFBN	40.00 / 71.11 / 52.44	90.00 / 71.11 / 82.44	50.00 / 88.89 / 65.56	50.00 / 88.89 / 65.56
APP	50.00 / 80.00 / 62.00	60.00 / 35.56 / 50.22	56.67 / 85.33 / 68.47	56.67 / 85.33 / 68.47
YGP	40.00 / 88.89 / 59.56	50.00 / 71.11 / 58.44	73.33 / 100.00 / 84.00	73.33 / 100.00 / 84.00
FAS	40.00 / 80.00 / 56.00	66.67 / 26.67 / 50.07	63.33 / 100.00 / 78.00	63.33 / 100.00 / 78.00

Analysis of Practicality

Practicality in assessment refers to the extent to which a test can be implemented efficiently, economically, and feasibly within a real educational setting (Brown, 2004). In this study, the practicality of the assessment instruments for the four English skills—writing, reading, listening, and speaking—was evaluated based on several factors: time efficiency, material accessibility, ease of administration, and student engagement during implementation. All tests were administered in regular class hours with no need for additional sessions or costly resources, which confirms their overall feasibility.

The assessments were distributed across four different class meetings, each focusing on one skill. Each session followed the school's 40–60-minute time slot. The writing test included multiple-choice questions related to grammar and vocabulary, as well as short essay prompts requiring sentence rearrangement or paragraph writing. Students completed the objective section within 15–20 minutes, allowing ample time for the essay section. Reading tasks involved multiple-choice and short-answer comprehension questions based on grade-level texts. The printed worksheets used were simple and clearly formatted, and students were able to follow instructions without difficulty.

For the listening test, the audio materials consisted of short pre-recorded dialogues or monologues of about 1–2 minutes each, followed by a series of comprehension questions. These were played through a smartphone connected to a classroom speaker, which functioned smoothly without the need for additional setup or external equipment. Since the teacher was familiar with the tools, the session proceeded without delays or technical issues.

The speaking assessment was conducted in a one-by-one format, where each student took turns standing in front of the class to introduce themselves in English. They stated personal details such as their name, age, hobbies, and favorite food—topics previously practiced during speaking lessons. This individual format allowed the teacher to assess pronunciation, fluency, vocabulary, and coherence in a controlled and manageable setting. The task required no additional tools, and students responded confidently, indicating their preparedness and comfort with the test format.

Across all four skills, students were able to complete their tasks independently and within the given time. The instructions were delivered clearly in English, and when needed, brief clarifications in Bahasa Indonesia ensured full comprehension. No students required excessive assistance, and there were no reports of the tests being too long or too difficult. Furthermore, all materials used were low-cost and readily available—such as printed worksheets, simple rubrics, and audio playback tools—making the assessments highly replicable in other similar school contexts.

In conclusion, the assessment procedures met the criteria of practicality in all evaluated aspects. The tests were easy to administer, aligned with the classroom schedule, did not rely on advanced technology or external support, and were well-suited to the students' proficiency level and learning environment.

Analysis of Reliability

The reliability analysis was conducted to assess the internal consistency of the test instruments for writing, reading, listening, and speaking skills, determining whether the set of items produced stable and consistent results across different students. For this purpose, the Cronbach's Alpha formula was applied:

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum\sigma_i^2}{\sigma_t^2}\right)$$

Where:

α = reliability coefficient

k = total number of items

$\sum\sigma_i^2$ = sum of the variances of individual items

σ_x^2 = variance of the total scores

The analysis for each section was performed using Microsoft Excel, where the variance for each item was computed using the =VAR.S (range) formula. The total score variance was computed from the final score column of all students.

Writing Test Reliability

For the writing test, the Cronbach's Alpha for the multiple-choice section (k = 30) was 0.86, with a sum of item variances of 6.12 and a total score variance of 29.3. For the essay section (k = 15), the Cronbach's Alpha was 0.843, with a sum of item variances of 19.36 and a total score variance of 82.94. According to Arikunto (2010) cited in (Adlin et al, 2021), a reliability coefficient of 0.70 or higher is considered acceptable, indicating that both sections demonstrated strong reliability, confirming the test's internal consistency.

Reading Test Reliability

For the reading test, the multiple-choice section (k = 30) yielded a Cronbach's Alpha of 0.59, indicating moderate reliability. With a sum of item variances of 5.68 and a total score variance of 12.7, this section's reliability was below the ideal threshold of 0.70, suggesting room for improvement in item construction. However, the essay section (k = 15) showed a significantly higher internal consistency with a Cronbach's Alpha of 0.85738, reflecting high reliability. The total item variance was 19.34, and the total score variance was 96.79, indicating that the essay items were highly consistent in measuring students' reading comprehension.

Listening Test Reliability

For the listening test, both the multiple-choice and essay sections demonstrated very high reliability. The multiple-choice section (k = 30) had a Cronbach's Alpha of 0.94, with a sum of item variances of 4.66 and a total score variance of 5.56. The essay section (k = 15) had a Cronbach's Alpha of 0.87678, with a sum of item variances of 19.13 and a total score variance of 105.29. These high reliability scores indicate that the listening test was consistently measuring the listening comprehension construct with excellent psychometric quality.

Speaking Test Reliability

For the speaking test, the Cronbach's Alpha was calculated using the same formula, yielding a value of 0.94. This indicates excellent reliability, as it exceeds the threshold for high internal consistency (Sugiyono, 2013). The calculated variance values were 4.63 for the sum of item variances and 49.6 for the total score variance, confirming that the speaking test is highly reliable in assessing speaking performance.

In conclusion, the tests for writing, reading, listening, and speaking skills all exhibited good to excellent internal consistency, with Cronbach's Alpha values well above the acceptable threshold of 0.70 for reliability. This demonstrates that the instruments used in the assessments are dependable and suitable for evaluating student performance across these skills, although some sections may benefit from further refinement.

Analysis of Validity

The analysis of validity in this study was carried out to determine how accurately the test items measured students' actual language abilities across the four language skills: writing, reading, listening, and speaking. The instrument used included a total of 45 items per skill: 30 multiple-choice (MC) questions and 15 essay questions. To assess the validity of each test item, the Pearson Product Moment correlation coefficient was applied, using the formula:

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

$$r = \frac{n\sum xX - (\Sigma X)(\sum Y)}{\sqrt{[n\sum x^2 - (\Sigma x)^2][n\Sigma Y^2 - (\sum Y)^2]}}$$

(Ary et al, 2014)

Where:

X is the score for each item,

Y is the total score for each student,

n is the number of students (n = 20 in this study).

In practice, the item scores were either binary (0 or 1) for multiple-choice questions, or based on a rating scale (e.g., 1–5) for essay responses. These individual item scores were correlated with the overall total test scores using Microsoft Excel's =PEARSON (range_item, range_total) function. The r-count (r-hitung) result for each item was then compared to the r-table value at a 5% significance level with 18 degrees of freedom (df = 20 - 2), which is 0.444. An item was classified as valid if r ≥ 0.444, and invalid if r < 0.444.

Table 2. Summary of Validity Analysis for Multiple-Choice and Essay Items

Skill	Total MC Items	Valid MC Items	Invalid MC Items	Total Essay Items	Valid Essay Items	Invalid Essay Items
Writing	30	13	17	15	11	4
Reading	30	17	13	15	15	0
Listening	30	29	1	15	15	0
Speaking	30	29	1	15	15	0

Based on the table above, the writing skill test demonstrated the lowest number of valid multiple-choice items, with only 13 out of 30 MC questions being valid, and 11 out of 15 essay questions meeting the criteria. This suggests that while a substantial portion of the essay questions effectively measured students' writing skills, the multiple-choice items may have failed to capture the complexity of writing ability or were not clearly aligned with the intended learning objectives.

In the reading skill test, the result was slightly better, with 17 MC questions found valid and 13 invalid, while all 15 essay questions were valid. This indicates that the essay section had strong content alignment and was well-constructed to measure students' reading comprehension, particularly in contextual and inferential understanding. However, the multiple-choice section still showed a need for revision due to the number of invalid items.

The listening skill test yielded highly positive results, with 29 valid multiple-choice questions out of 30, and only one item considered invalid. This suggests that the test items for listening were highly aligned with the listening skills being measured, including the ability to understand spoken information, identify details, and interpret audio-based content.

Similarly, the speaking skill test also demonstrated strong validity, with 29 out of 30 MC items being valid and all 15 essay questions validated. These findings imply that the test for speaking was well-constructed and appropriately measured students' oral communication abilities, both through their knowledge of speaking-related expressions (via MC questions) and their actual spoken performance (via essay prompts).

From the results, it is evident that essay questions tend to exhibit higher validity across all skills, likely due to their open-ended nature, which allows students to demonstrate deeper understanding and actual language use. On the other hand, multiple-choice items, especially in writing and reading, showed a greater number of invalid items. This might be due to poorly constructed distractors, vague options, or a lack of alignment with the cognitive demands of the skill being assessed.

Overall, this validity analysis indicates that while the assessments for listening and speaking are highly effective in measuring student performance, the writing and reading tests—particularly their multiple-choice sections—require careful revision and improvement to enhance their measurement accuracy and ensure alignment with learning objectives. These improvements would help maintain the integrity of the language

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

assessment and provide more accurate evaluations of students' language skills.

Analysis of Authenticity

In this study, authenticity in the language assessments was evaluated by examining the alignment of task types, themes, and students' response patterns with real-world language use. The assessment tasks were designed to reflect authentic communicative situations that students might encounter outside the classroom. Listening tasks covered topics such as daily activities, shopping, personal introductions, and school life, while speaking tasks required students to introduce themselves and describe hobbies or memorable experiences. Reading passages were drawn from familiar topics, and writing prompts encouraged personal storytelling, all aimed at creating a realistic and relatable experience for the students.

Students' performance on these tasks further demonstrated the authenticity of the assessments. In listening, almost all students scored above 60, with some reaching 90 or higher, suggesting that they were comfortable with the listening format and content. For example, Arief Syugi Sireger (92.07), Revand (99.47), and Yesaya (98.67) excelled, indicating familiarity with the content and format. Similarly, in reading, students like Muhammad Farhan Maulana (96.00) and Arief (86.13) scored high, showing ease with authentic reading passages.

However, productive skills like writing and speaking showed more variation in scores. Writing scores ranged from 18.67 (Mikael F. Sinaga) to 80.27 (Fikri D. Ramadhan), while speaking scores ranged from 41.78 to 99.47. While some students struggled with these skills, the personal and relatable nature of the tasks ensured that the difficulties were due to language ability, not unfamiliarity with the task type. This further supports the authenticity of the assessment process, as real communication often requires more effort than recognition-based tasks.

The integration of both multiple-choice and essay formats also contributed to authenticity. Multiple-choice questions mirrored real-life communication functions, such as identifying correct phrases in a conversation, while essay tasks required students to generate language, organize ideas, and speak or write meaningfully. These tasks encouraged students to engage in spontaneous, real-world language use, especially in speaking tasks where they responded to unrehearsed questions based on their own lives.

Despite the fact that essay scores were generally lower than multiple-choice scores, this reflects the authenticity of the assessment process. Producing language (as required in essay tasks) is inherently more cognitively demanding than recognizing correct answers. The average differences between the multiple-choice and essay sections across skills were as follows:

Table 3. Comparison of Average Multiple-Choice and Essay Scores Across Skills

Skill	Average MC Score	Average Essay Score	Difference
Writing	54.77	79.91	25.14
Reading	73.67	78.67	5
Listening	75.08	94.93	19.85
Speaking	75.08	94.93	19.85

These findings are consistent with Astuti and Prasetyo (2024), who reported that authentic tasks—such as personal storytelling, real-life dialogues, and contextualized comprehension—tend to produce more meaningful engagement and a more accurate representation of language proficiency. Similarly, Giraldo (2022) emphasized that authenticity in assessment fosters learner motivation and encourages deeper cognitive processing, which in turn leads to better performance in communicative tasks. In line with these studies, the present findings suggest that relatable, real-world tasks in both receptive and productive formats not only reflect actual language use but also promote student willingness to communicate and perform.

Analysis of Washback

The study analyzed the washback effect of assessments on student engagement, performance, and instructional responses. The findings showed that the testing process had a positive impact on student motivation and learning, particularly in speaking and listening tasks. Students demonstrated increased enthusiasm and engagement, with many achieving higher scores in essay-based and oral tasks compared to

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

multiple-choice (MC) tasks. Here is the summary of the average scores:

Table 4. Average Scores in Multiple-Choice and Essay-Based Assessments

Skill	Average MC Score	Average Essay Score	Difference (Essay−MC)
Writing	55	82.2	27.2
Reading	74.67	76.51	1.84
Listening	74.17	93.47	19.3
Speaking	74.17	93.47	19.3

These findings align with prior research by Pratiwi and Nugroho (2023), who reported that authentic, performance-based assessments often generate stronger positive washback than traditional recognition-focused tests. Similarly, Giraldo (2022) observed that productive tasks foster deeper cognitive engagement and promote meaningful language use, resulting in improved performance. The present study reinforces these conclusions, showing that assessments emphasizing authentic and communicative tasks can enhance learner motivation and outcomes. Additionally, the performance patterns observed here highlight areas for instructional improvement, such as providing more targeted support for writing coherence and paragraph organization.

CONCLUSION

Based on the analysis of English assessment instruments used for Grade VII students, this study found that the tests generally fulfilled the five principles of language assessment: practicality, reliability, validity, authenticity, and washback. The validity analysis using the Pearson Product–Moment correlation indicated that most essay items in speaking and writing were valid, while some multiple-choice items in reading and writing required revision. Reliability results using Cronbach's Alpha showed excellent reliability for listening (0.94) and speaking (0.94), high for writing (0.86), and moderate for reading (0.59), suggesting that the reading multiple-choice items were less consistent. Practicality was achieved as all tests were administered effectively using available classroom resources and time, while authenticity was evident in tasks that reflected real-life communication, such as describing daily routines or writing short messages. Positive washback was also observed, as the assessments increased student engagement and provided meaningful feedback for both learners and teachers. Overall, the instruments effectively measured students' language proficiency—particularly in listening and speaking—and supported instructional improvement, although certain aspects, especially in the reading section, require further refinement.

SUGGESTION

Teachers are encouraged to revise multiple-choice questions—especially in reading—to improve clarity and reliability. Essay and oral tasks should be used more frequently to promote authentic language use. Schools should facilitate training and collaboration for better assessment design. Future researchers may expand the study scope or integrate digital assessment tools to enhance effectiveness and scalability.

References

Abrar-ul-Hassan, S., & Nassaji, H. (2024). *Language assessment and teacher reflection: A systems view*. Routledge.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129.

Adlin, D., Situmorang, J., & Saragih, A. H. (2021, November). The Development of Assessment Instrument Models of Malay Dance Art. In *6th Annual International Seminar on Transformative Education and Educational Leadership (AISTEEL 2021)* (pp. 584-591). Atlantis Press.

Alqahtani, F. (2021). The Impact of Language Testing Washback in Promoting Teaching and Learning Processes: A Theoretical Review. *English Language Teaching, 14*(7), 21-26.

Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* (2nd ed.). Bumi Aksara.

Astuti, Y. D., & Prasetyo, T. R. (2024). Designing authentic assessment in EFL classrooms: Challenges and strategies. *Journal of Language Teaching and Research, 15*(1), 45–53.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*.

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

Oxford University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.

Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (2nd ed.). Pearson Education.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.

Gay, L. R., Mills, G. E., & Airasian, P. (2012). *Educational research: Competencies for analysis and applications* (10th ed.). Pearson Education.

Giraldo, F. (2022). Language assessment literacy for pre-service and in-service teachers. *Language Testing, 39*(2), 251–276.

Handayani, L., & Susanti, R. (2022). Practical language testing in Indonesian EFL classrooms. *Jurnal Pendidikan Bahasa*, *17*(2), 89–97.

Hartati, R., & Siahaan, A. (2023). Assessing speaking performance in EFL settings: Challenges and solutions. *Journal of Applied Linguistics and Language Research, 10*(1), 33–42.

Jabbar, A. (2023). The role of technology in post-pandemic language assessment. *Asian Journal of Education and e-Learning, 11*(2), 23–29.

Kang, S., & Wu, Y. (2021). Designing effective listening assessments in EFL classrooms. *Language Testing in Asia, 11*(1), 1–12.

Lee, J., & Butler, Y. G. (2020). Developing language assessment literacy: A framework for teacher education. In S. Abrar-ul-Hassan & H. Nassaji (Eds.), *Perspectives on language assessment literacy* (pp. 11–30). Routledge.

Lestari, I., & Hamdani, A. (2020). Assessing students' writing through authentic tasks. *Journal of English Education Studies, 3*(1), 56–64.

Marzuki, A., & Rukmini, D. (2022). Assessing reading comprehension: A focus on inferencing and digital literacy. *Indonesian Journal of Language Assessment, 4*(2), 89–97.

Masrur, F., & Kuning, D. S. (2025). The Correlation Between Self Confidence and Speaking About DescribingbThing of the Eleventh Grade Student of SMK Negeri 1 Abung Selatan Academic Year 2023/2024. *Griya Cendikia*, *10*(1), 142-159.

Pratiwi, E. Y., & Nugroho, D. (2023). Positive washback in English language assessment: A case study in Indonesia. *ELT Research Journal, 12*(1), 14–26.

Pusey, K. (2024). Interrogating the Concept of Authenticity in Language Testing: A Narrative Review.

Riazi, A. M., & Farsani, M. A. (2024). Mixed-methods research in applied linguistics: Charting the progress through the second decade of the twenty-first century. *Language Teaching*, *57*(2), 143-182.

Saputra, M. R., & Fitriani, R. (2023). Reliability in language testing: Theory and classroom application. *Jurnal Ilmu Pendidikan, 9*(1), 1–10.

Sianturi, L. P. D., Silalahi, D. E., & Purba, C. N. (2020). Improving students' writing ability based on higher order thinking skills (HOTS) questions at 8th grade in SMP Swasta Kartika 1-4 Pematangsiantar. *JETAFL (Journal of English Teaching as a Foreign Language)*, *6*(2), 1-17.

Sihombing, B. A., Sinurat, B., & Silalahi, D. E. (2024). The Effect of The Jigsaw Teaching Technique on The Students' Reading Comprehension at Seventh Grade of UPTD SMPN 5 Pematangsiantar. *TOFEDU: The Future of Education Journal*, *3*(5), 1631-1639.

Silalahi, D. E. (2017). The Effect of Small Group Discussion Technique on Students Reading Comprehension Ability.

Sugiyono. (2017). *Metode penelitian pendidikan: Pendekatan kuantitatif, kualitatif, dan R&D*. Alfabeta.

ISSN: 2088-0316 | e-ISSN: 2685-0133
https://doi.org/10.37630/jpb.v15i3.3331

Suryanto, J. (2025). Washback effect: How it works on EFL students. *Celtic: A Journal of Culture, English Language Teaching, Literature and Linguistics*, *12*(1), 323-339.

Wibowo, H., Yauri, A. M., & Hasanah, U. (2021). Using whatsapp as a learning media in teaching writing. *International Journal of Research on English Teaching and Applied Linguistics*, *2*(1), 72-79.

Wijaya, R., & Salim, M. (2021). Validity and reliability in designing English reading assessments. *Journal of English Teaching and Learning, 10*(2), 87–95.

Winna, D. A., & Sabarun. (2023). Integrating standardized and classroom-based assessment: A balanced approach. *Indonesian Journal of English Education, 10*(1), 30–44.