

ANALISIS PRINSIP PENILAIAN DAN EVALUASI BAHASA MELALUI TEST BAHASA INGGRIS

Rahel Sastri Rodearnita Purba

rahelsastri03@gmail.com

Universitas HKBP Nommensen Pematangsiantar

Riby Astri Siboro

ribyastri7@gmail.com

Universitas HKBP Nommensen Pematangsiantar

Hery Renaldi Tarigan

herytarigann@gmail.com

Universitas HKBP Nommensen Pematangsiantar

Rani Puspa Sari

ranipuspa566@gmail.com

Universitas HKBP Nommensen Pematangsiantar

Dumaris E Silalahi

dumaris.silalahi@uhn.ac.id

Universitas HKBP Nommensen Pematangsiantar

ARTICLE INFO Article

History: Received Date: 10th

Oct 2025 Received in

Revised Form Date: 27th Oct

2024 Accepted Date: 29th

Oct 2025 Published online

Date 31th Oct 2025

Keyword: Analyses,
Language, Assessment,
Evaluation, English tests

This study analyzes the implementation of English language assessment based on the five fundamental principles of evaluation: practicality, reliability, validity, authenticity, and washback. The assessment was designed to evaluate students' competencies in listening, speaking, reading, and writing through a combination of multiple-choice and essay questions. The test was administered to 28 eight-grade students at SMP Taman siswa Pematangsiantar. Employing a descriptive qualitative approach with quantitative support, the research incorporated inter-rater reliability checks and item difficulty analysis to ensure comprehensive evaluation. The findings reveal that the English test was practical within the constraints of real classroom settings, demonstrated high reliability through consistent scoring across raters, and achieved validity by effectively measuring the targeted language skills. Furthermore, the test tasks reflected authentic, real-life language use and generated positive washback by enhancing teaching practices and learning outcomes. Notably, the results highlighted the significance of grammatical accuracy, particularly in productive skills such as speaking and writing. This study offers a practical model for implementing the principles of language assessment in Indonesian EFL classrooms, even with limited resources, and serves as a useful reference for educators and researchers in developing valid, reliable, and pedagogically sound assessment tools

PENDAHULUAN

Assessment plays as a vital role in evaluate student's contribution as participant to analyze their result of answering multiple-choice and essay test into five principles of language assessment and evaluation, especially in English language learning. In junior high schools of Taman Siswa, assessment is commonly administered through multiple-choice and essay questions that aim to measure students result test based on practicality, validity, reliability, authenticity and washback. This method of evaluation considered to all language skills that differently give conceptualization of design test to recognize each function of Listening, speaking, reading and writing test based that are divided into receptive skills listening and reading and productive skills speaking and writing each involving either oral or written models of communication. Understanding the nature of these skills is crucial for both instruction and assessment, as each requires distinct strategies and levels of cognitive engagement.

Page | 36

However, in many classroom practices, these skills were often to emphasize grammar memorization, and vocabulary lists rather than communicative competence. This imbalance limits the opportunity for students to demonstrate their actual language use in real-life contexts. Therefore, meaningful assessment must reflect both the instructional goals and real-world application of language. To address this both space of range, language educators must refer to clear assessment frameworks that guide test construction and implementation effectively.

Brown (2004), in his book *Language Assessment: Principles and Classroom Practices*, emphasizes that a good assessment must reflect real-world language use and provide feedback that helps learners improve their performance. This is achievable only when assessments are clearly linked to the specific skills being taught. In line with this, Richards and Renandya (2002) argue that language learning becomes more effective when all four language skills are developed in an integrated manner and when assessment strategies support this integration. These perspectives show that meaningful assessment goes beyond scoring; it directly influences learning outcomes and instructional quality.

In the context of Indonesian education, especially in classrooms with limited facilities, implementing assessments based on such principles presents both challenges and

opportunities. The reality on the ground shows that many students still face difficulties in understanding of English sentence applying vocabulary, expressing ideas orally, or understanding spoken and written texts. Besides that, we implement this test through approaches to the students' participants at eight grade classrooms. This observation raises several concerns regarding how English assessments are designed and whether they align with students' actual abilities and classroom realities. Teachers often find it difficult to construct valid tests that address the four language skills in a balanced way and still meet curriculum requirements.

This study was conducted to examine how English tests can be developed and implemented based on five fundamental principles of assessment: practicality, reliability, validity, authenticity, and washback, as outlined by Brown. The research focuses on eighth-grade students at SMP Taman Siswa Pematangsiantar or somehow called as Taman Dewasa. Then, analyzes their performance in listening, speaking, reading, and writing through structured test items. These tests included multiple-choice questions for receptive skills and essay tasks for productive skills. The results were used to observe how well students could apply their knowledge and to what extent the test reflected real communicative demands.

The core aim of this study is to analyze the extent to which language assessment and evaluation principles are applied in real classroom testing. It also seeks to identify the strengths and weaknesses in students' performance across the four skills, and to understand the effectiveness of the test items used. Through this analysis, the research offers practical insights that can support English teachers in constructing better assessments. Moreover, the findings are expected to contribute to the development of fair, valid, and reliable assessment strategies that support language learning in Indonesian EFL classrooms, particularly in schools where resources and teaching support may be limited.

This study employed a descriptive quantitative method supported by qualitative result. Research design aimed to assess students' English language abilities based on the four language skills listening, speaking, reading, and writing. The assessments were conducted using multiple-choice and essay tests, and the results were analyzed through

the lens of Brown's (2004) five principles of language assessment: practicality, reliability, validity, authenticity, and washback.

The research participants were collected from 28 students in VIII-B classrooms at Taman Siswa Junior High School or usually called as Taman Dewasa willing to help us in the free time where at the time they're just completed final examination to reach the next grade, and they were participated started on 18 June 2025 at 7:30 a.m.

Page | 38

Data collection was conducted through structured testing instruments composed of 30 multiple-choice questions and 15 essay questions, all of which were distributed across the four language skills. On the first day, students were asked to answer paper sheets about speaking at 7:30 a.m. and finished to 11.00 o'clock and because friend and I were in the same class at the time, so we conduct a session it means we provide two skills for a day. For the next day on Thursday or second day, they completed tasks focused on listening with the same technique after part one of session in the rest of time my friend Riby have controlled to the class. In the next day on Friday same as usual I put the test onto Reading and writing for a day and the rest of it. In calculated taking three days off to process the examination complete. Basically, students at Tamansiswa really have a big deal to complete all the task because their very weak ability in English Language.

The students' test results were collected and analyzed using descriptive statistical techniques, including percentage and average scoring. Each student's performance in the four language skills was categorized based on achievement levels. In addition to the quantitative analysis, a qualitative review was conducted to evaluate the quality of the test implementation in accordance with the five principles of language assessment. This involved examining the practicality of test administration, the consistency and objectivity of scoring, the validity of the test items in measuring intended skills, the authenticity of the tasks in reflecting real-life language use, and the washback effect or influence of the assessments on students' motivation and language learning outcomes. This method thus provided a comprehensive evaluation, combining both performance data and the pedagogical value of the assessment instruments within the context of junior high school English instruction. Let's break it down into fifth principle

METODE

Analyzing fourth language skills tests into fifth principle, here the structure explanation.

A. Practicality

Practicality defined as the relationship between available resources for the test, human resources, material resources, time resources and resources which will be required in the design development and use on the test (Bachman & Palmer 1996:35-36)

According to Brown (2004:19) describe practicality as cost, time, administration and scoring or evaluation.

1. Cost

Cost as term define that the test should not be too expensive to conduct, the cost of the test must stay within the budget and avoid conducting a test that requires excessive budget.

Noted: Students didn't have to pay for the tests; meanwhile, we, as the responsibility for the test should prepare printing of paper sheets.

2. Time

Time accessibility to test students begun in the first day on Wednesday until to Friday, it means took three days off to finish the process of tests. For the specific time start at 7:30-10.00 a.m. in this case students took two hours thirty minutes to complete speaking test.

Analysis from time management:

$150 \text{ minutes} \div 45 \text{ questions} = \pm 3,3 \text{ minutes per question}$

Multiple choice (30 questions): reasonable to give time 1,5–2 minutes/questions → required $\pm 45\text{--}60$ minutes

Essay test (15 questions: because the students low average in English subject, reasonable to give students time 4–5 minutes/questions → required $\pm 60\text{--}75$ minutes

The properly time: 45 to 60 minutes (MCQ) + 60 to 75 (essay) = 105 to 135 minutes or 1 hours 45 minutes – 2 hours 15 minutes so, totally fine took 2 hours 30 minute less and consistently reasonable , even it have more a space for students to ask something from test that cannot be understood, and help students not to rush.

3. Administration

In administration, this term was implicating the qualification of test simply relate to every structure question and not to be complicated or complex to conduct in this administration preparation test class in order, which mean student have rights to feel comfortable in

process of study. So, at this test project we let student have free will to speak up and asking.

Illustration: Students in lower skill of knowledge in language competency have no ideas to answer the question and end up asked the teacher. So, at this point student hardly to found out the answer and waste time or break management in time process, for this solution as a guidance student might be guide and direct to understand more deeply about the functional of language skill base and have curiosity to learn.

4. Scoring and evaluation

Scoring and evaluation process should be fits into the time and allocation. Especially, to evaluate multiple choice and essays tests into result where to make it easy and clear in time and allocation sessions. Assessor as the responsive, in this stage to scoring test one by one of each skills test should be accompanied with scoring rubrics and key answer so it can be effective in time management. The scoring pattern method for MCQ every student who answer the right answer completely gained 30 points, it means one question for one point, while essay tests session was collected based on 1 to 4 range of classification answer in order into the holistic rubric or assess all the answer in essays part. Listening was evaluating students how they're receive information based on audio from You Tube and script description from book in Merdeka curriculum for class viii-b. Speaking was conducted to answer the questions based on the dialogue at paper sheets, how they communication and mentioned every word to pronounce correctly but most of them very hardly to speak in English. In this case to let them try their best it's influence their self-confidence. In reading and writing evaluate investigating that students very weak at this skill because instead of reading and write the questions they're just give up so easily.

B. Validity

Validity of a test is the extent on capability of students, which is to measures what it is supposed to measure (Huges, 2003:26).

According to Heaton 1990:159 validity refers to a test's ability to accurately measure what it is intended to measure, and nothing else. Essentially, a valid test ensures that the results truly reflect the specific language skills or knowledge being assessed.

There are several parts in principle validity based on brown perception.

1. Content validity describes the correlation between the contents of the test and the language skill, structurally in main point.

In listening skill, we as assessor provide speaker as media tool to help students to analyze every information from video and somehow, if they're hard to understand I speak loudly to make sure it was clearly.

In speaking we provide dialogue script to make students perform one by one in front of the class, even there are most of them rejected the challenges.

Reading and writing skill in the context used self-conscious to answer the questions properly based on text in the passage like descriptive text, fable stories etc.

2. Criterion validity, the test was analyzed based on the **criterion validity** principle to see whether the test results matched other trusted sources of student performance. In this case, students' scores from the English language skills test (listening, speaking, reading, and writing) were compared with their English subject grades in the school report.

The result showed that most students who got high scores in the test also had good grades in English class. On the other hand, students who performed poorly in the test also had low classroom performance, especially in writing and speaking activities. This shows that the test had a clear relationship with the students' actual English abilities in class.

Therefore, the test demonstrated **good criterion validity**, because the scores reflected the students' real performance and matched the teacher's observation and school report.

3. Construct validity in this test aimed to assess students' English skills across four areas: listening, speaking, reading, and writing. However, many students struggled to understand the test instructions and questions, which were written fully in English. Because of that, it became unclear whether the test was really measuring their actual language skills or just their ability to understand the test format and vocabulary.

For example, the speaking section asked students to choose or write expressions in writing, not to speak. Similarly, students found the writing tasks confusing, even though some of them could write better when guided in class. These problems suggest that the test may not have fully captured the true construct of each skill — especially for students with low proficiency.

In short, the test had **limited construct validity** because it didn't always reflect the real abilities students used in the classroom. To improve this, future tests should use simpler

language, clearer instructions, and tasks that truly represent each skill in a more practical and familiar way

C. Reliability

To determine the reliability of their English language test instruments, Rahel and Riby employed the parallel-form method, also known as alternate-form reliability, rooted in classical test theory and supported by Thorndike and Hagen (1977). This method evaluates consistency between two equivalent test forms administered under similar conditions.

Each researcher developed a separate test measuring the same language skills—listening, speaking, reading, and writing—with equivalent difficulty and structure. The tests were given to the same junior high school students within a short time interval to control learning and fatigue effects.

To ensure objectivity, Rahel scored Riby's test, and Riby scored Rahel, avoiding personal bias. A structured scoring system I used: 1 point for each correction in multiple-choice answer; 1-4 points range for correction for 1 to 15 essay test questions. Scoring was guided by detailed rubrics focusing on task completion, coherence, grammar, vocabulary, and relevance. Besides that, Riby scored the students for multiple-choice one points for each answer and for essay 1 to 5 questions value six points, and then six to fifth teen questions values four points.

A correlation analysis was conducted to assess reliability. A high correlation between the two test forms confirmed that both produced consistent results. The involvement of two raters and cross-scoring enhanced credibility, especially for subjective responses in speaking and writing.

In conclusion, the rigorous application of the parallel-form method by Rahel and Riby supported by theoretical and empirical practices, ensured the reliability, objectivity, and fairness of their English language assessments.

Analysis: In the table1. Result of reliability principle, many students experienced large differences between their scores on the two versions of the speaking test. Only a few students, such as Arya and Stevan, showed consistent performance. The majority, however, had inconsistencies above 25 points, indicating that the test items may not have had stable levels of difficulty. This suggests that the reliability of the speaking test is still

low to moderate and would benefit from further item revision to ensure fairer measurement.

D. Authenticity

Authenticity defined as the real world, which means is the degree of correspondence of the characteristics of a given language test task to the features of a target language task (Brown 2004:28)

Page | 43

The authenticity of the test was analyzed by reviewing how closely the tasks reflect real-life language use. The listening section included a dialogue about celebrating Independence Day with traditional games, which is highly contextual and familiar to students. The speaking test stimulated a real conversation between friends about favorite sports. In reading, the texts included a hygiene procedure and a moral fable both relevant to students' daily lives. Writing tasks asked students to describe their favorite food and share personal preferences, which are also realistic tasks. Based on this analysis, the test demonstrates a strong level of authenticity across all four skills.

The fourth language test were collected from real world question, for example:

1. Listening contains a dialogue from audio between father and son who celebrated Independence Day. Galang as a son and Pak Rahmansyah as Galang's father and they're talking about a lot of games included Panjat Pinang and there are many prizes at the top of the tree.
2. Speaking contains a real-world dialogue in real life situation at canteen, two friends who's talking about their favorite hobbies, invitation and gratitude situation.
3. Reading contains procedure text about "let's wash our hands properly" there are step by step to wash hands properly. For the rest of questions provide Fabel text about "the smart Deer and the Crocodile."

E. Washback

Washback as principle of language assessment and evaluation is commonly used in applied linguistics. It is rarely found in dictionaries. According to brown define washback or backwash refers to influence of testing on teaching and learning. There are two side of washback it can influence as positive side and negative side.

1. Positive washback has beneficial influence on teaching and learning. It means teachers and students have a positive attitude toward the examination or test and work willingly and collaboratively towards its objective.
2. Negative washback does not give any beneficial influence on teaching and learning. Tests which have negative washback is considered to have negative influence on teaching and learning

The analysis of washback in this study focused on the effects that the assessment had on students' learning attitudes, classroom performance, and instructional practices. Based on classroom observations and student outcomes, the assessment of four English language skills (listening, speaking, reading, and writing) produced both positive and negative washback effects.

On the positive side, the test successfully encouraged students and teachers to engage with all four core language skills, especially speaking and writing, which are often neglected in regular classroom instruction. The use of contextual materials such as dialogues, fables, and procedural texts helped expose students to real-life language use, enhancing the authenticity of the learning experience. Additionally, the inclusion of both multiple-choice and essay formats prompted the use of rubrics and more objective scoring, providing students with clearer feedback and helping teachers evaluate student performance more systematically.

However, the study also revealed several negative washback effects. Many students struggled with full-English instructions, which limited their ability to understand and respond to questions not due to a lack of language skill, but because of unfamiliar wording and complex formats. This led to low scores, especially in speaking and writing, and created frustration and anxiety among students. The lengthy duration of the test (2 hours 30 minutes per skill) further reduced student focus and motivation, particularly for those with lower proficiency. As a result, some students became disengaged and lost confidence in their language learning abilities

HASIL DAN PEMBAHASAN

The results of this study show that students' performance in the assessment of four English language skills listening, speaking, reading, and writing was generally below the expected standard. Out of 28 participants, the majority scored low, especially in the speaking and writing sections. Most students had difficulty understanding the test instructions and required additional guidance during the test.

Page | 45

Each skill was assessed through a combination of 30 multiple-choice questions and 15 essay items, totaling 180 questions across all four skills. The highest scores were generally found in listening and reading, while productive skills such as speaking and writing resulted in lower performance. Students' unfamiliarity with full-English question formats and lack of exposure to open-ended tasks contributed to the low achievement levels.

The test results were further analyzed based on Brown's five principles of language assessment:

a. Validity

The test demonstrated good content validity, as it was based on the English curriculum and covered a wide range of relevant tasks, including dialogues, fables, and procedural texts. However, there were issues with construct validity, especially in speaking and writing. Many students failed to perform not because they lacked the skill, but because they did not understand the instructions or question format. This indicates a gap between what the test intended to measure and what it assessed.

b. Reliability

The reliability of the test was found to be moderate to low. This was evident in the score inconsistency between two versions of the test given to the same students. Some students showed large score differences when facing different versions of speaking or writing tasks. This suggests that some items may not have had consistent difficulty levels, reducing the reliability of the test outcomes.

Four language Skills	Average score for all the students	High score	Low score
Listening	57,05	69	22
Speaking	68,00	87	58
Reading	45,11	82	0
Writing	71,26	89	50

Table 1. show the Riby's assessment method result

Categories	Score	Students' achievement	
Excellent	301-360	0	0%
Good	241-300	0	0%
Fair	181-240	1	3,6%
Poor	121-180	14	50%
Very poor	1-120	13	46,4%

Table 2. show the difference assessment method from Riby's to my results

The reliability result test from two difference way shows the why the assess conducted not in the sequence, at this point showing unrelated perspective of students' result test. Overall, it is because the different implementation, paper sheets questions and the comparison among student's whom available in present, when Riby's part only 19 students come along compared to me that having 27 students' presence for speaking skills, Listening were 24 students, reading and writing 20 students who contribute as participants.

Information added: Among the 19 students present in both tests, only 7 students (36.8%) demonstrated consistent performance across both assessments. The remaining 12 students (63.2%) showed significant score fluctuations, particularly in Reading and Speaking.

For example: Fahri Anggara scored low in Reading in the first test (24/90) but achieved 60/90 in the second test. Stevan Lim scored 71/90 in Reading (test 1) but only 60/90 in Writing (test 2). This indicates a lack of score stability, suggesting weak reliability between the two assessments.

c. Authenticity

Most tasks demonstrated high authenticity. The materials used in the test such as dialogues about daily situations, traditional events (e.g., Independence Day), fables, and food descriptions closely mirrored real-life communication that students might encounter. This helped create a more meaningful assessment context, despite their low proficiency.

d. Practicality

In terms of practicality, the test was manageable in one classroom with 28 students and did not require advanced technology. However, the test was long and dense each skill took up to 2 hours and 30 minutes, which was exhausting for students. Teachers also needed significant time and effort to grade the essay sections. Simpler rubrics and shortened tests could improve practicality.

e. Washback

The test had both positive and negative washback. Positively, it encouraged students and teachers to focus on all four skills and introduced varied, real-world-based tasks. However, the use of full-English instructions caused confusion, and many students became demotivated due to poor performance. Some students disengaged during the test, indicating a negative psychological effect

PENUTUP

In the conclusion this study concluded that the English language proficiency of students at SMP Taman Siswa Pematangsiantar across the four language skills listening, speaking, reading, and writing was generally below the expected level. This was particularly evident in the productive skills (speaking and writing), where students struggled to perform well. The main difficulties stemmed from students' limited understanding of instructions written entirely in English and their unfamiliarity with essay-based and situational tasks.

Page | 48

Based on the analysis of the five language assessment principles, the following findings were drawn: Content validity was strong, as the test aligned with the curriculum and involved contextual materials. However, construct validity was weak, as some items did not fully measure the intended skills. Reliability was found to be moderate to low, due to inconsistency in student scores across different test versions. Authenticity was high, since the test tasks reflected real-life contexts relevant to the students. From the practicality perspective, the test was manageable, but the scoring process especially for essay items required considerable time and effort. The washback effect was mixed: positively, the test promoted attention to all four skills, but negatively, it led to stress and demotivation among students due to the test's length and complexity. To improve the quality of assessment in future applications, the following suggestions are proposed: Test instructions should be written in simpler English or delivered in bilingual format to support student understanding. Test formats should be adjusted to better match students' actual abilities and familiarity, especially in speaking and writing. Essay assessments should use clear and consistent scoring rubrics, and the structure should be simplified to make correction more efficient. The number of items and test duration should be optimized to avoid student fatigue, and test sessions may be divided if necessary. Teachers should be trained to develop test items that are valid, authentic, and aligned with classroom learning, to maximize the positive impact of assessment on student learning

REFERENSI

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Page | 49

Anderson, J. (2005). *Developing tests and assessments for teaching English*. British Council.

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

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

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Green, A. (2007). *Washback in language testing: Research contexts and methods*. Cambridge University Press.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2017). *Kurikulum 2013: Kompetensi inti dan kompetensi dasar SMP/MTS*. Kemdikbud.

McNamara, T. (2000). *Language testing*. Oxford University Press.