INNOVATIVE: Journal Of Social Science Research
Volume 5 Nomor 4 Tahun 2025 Page 4711-4722
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Effectiveness Of Language Assessment Tests: An Observation Based On Reliability, Validity, Authenticity, Practicality, And Washback Principles In Class 7a At SMP Negeri 4 Pematangsiantar

Dhea Natasya Sitepu[1✉], Sonya Lauri Situmeang[2], Ribby Violin Sembiring[3],
Dyvia Chayani Saragih[4], Dumaris E. Silalahi[5]
Universitas HKBP Nommensen Pematangsiantar
Email: dumaris.silalahi@uhn.ac.id[1✉]

Abstrak

Laporan observasi ini meneliti efektivitas butir-butir tes bahasa Inggris yang digunakan di Kelas 7A SMP Negeri 4 Pematangsiantar. Studi ini mengevaluasi tes berdasarkan prinsip-prinsip utama penilaian bahasa, termasuk validitas, reliabilitas, kepraktisan, keaslian, dan umpan balik. Data dikumpulkan melalui observasi langsung selama pemberian tes pilihan ganda dan esai, yang menilai keterampilan siswa dalam menyimak, berbicara, membaca, dan menulis. Analisis kuantitatif dilakukan dengan menggunakan SPSS 24 untuk mengukur reliabilitas dan validitas dari butir-butir tes. Temuan menunjukkan bahwa tes pilihan ganda menunjukkan reliabilitas yang tinggi, sementara tes esai menunjukkan konsistensi yang moderat, yang menunjukkan perlunya perbaikan rubrik penilaian. Tes ini terbukti praktis dan otentik, dengan efek washback yang positif terhadap motivasi dan hasil belajar siswa. Rekomendasi diberikan untuk meningkatkan desain tes di masa depan untuk memastikan keselarasan dengan tujuan kurikulum dan penggunaan bahasa siswa di dunia nyata.

Kata Kunci: *Tes bahasa Inggris, penilaian bahasa, prinsip penilaian bahasa*

Abstract

This observation report examines the effectiveness of English test items used in Class 7A at SMP Negeri 4 Pematangsiantar. The study evaluates the tests based on key principles of language assessment, including validity, reliability, practicality, authenticity, and washback. Data were collected through direct observation during the administration of multiple-choice and essay tests, which assessed students' skills in listening, speaking, reading, and writing. Quantitative analysis was conducted using SPSS 24 to measure the reliability and validity of the test items. Findings indicate that the multiple-choice tests demonstrated high reliability, while essay tests showed moderate consistency, suggesting a need for improved scoring rubrics. The test was found to be practical and authentic, with a positive washback effect on student motivation and learning outcomes. Recommendations are provided to enhance future test design to ensure alignment with curriculum goals and students' real-world language use.
Keywords: *English Test, Language Assessment, Principle of Language Assessment*

INTRODUCTION

In the learning process, evaluation is considered an inseparable part (Fiska et al., 2021). Evaluation consists of two main processes: measurement, which involves comparing something being examined using systematically organized tools, and assessment, which refers to the interpretation of the results from those measurements (Elviana, 2020; Prasad, 2024; Fitriani & Rahmadewi, 2025). Evaluation is regarded as important because it helps assess the extent to which learning objectives have been achieved, and its results can serve as a basis for further follow-up actions (Lesta Ariany et al., 2018; Abdurrasyid & Panggabean, 2024; Retno et al, 2025). Through evaluation, the quality of teaching and learning is expected to improve continuously, which in turn influences the development of students' skills. The improvement of students' abilities contributes to the overall enhancement of education quality, as educational quality is aligned with students' capabilities (Khasanah et al., 2023).

Based on the role of tests as measurement tools, the success of a test can be evaluated by its ability to provide information that accurately reflects the actual condition of the object being measured. Test is a method, a tool or an instrument for measuring students' ability, mastery, or achievement of learning (Brown & Abeywickrama, 2010:3; Puspita & Syihabuddin, 2020; Hayati et al, 2020). The tool or instrument here can be in the forms of questions to be answered by students, true-false items or multiple choice items for students to answer. The next term is assessment which is claimed to have a wider meaning. It includes formal tests and also informal tests. Informal tests are usually incidental or unplanned, and can be in the forms of observation and/or comment (Brown & Abeywickrama, 2010: 3). Therefore, before being applied, a learning outcome test needs to be thoroughly examined to ensure it can produce optimal results (Aulia Ulfa Dewi, 2016).

Evaluation plays a vital role in the learning process as it helps determine whether learning objectives have been achieved. One of the most commonly used forms of evaluation in schools is testing, particularly multiple-choice tests. These tests must be carefully constructed using appropriate assessment principles to ensure they accurately and objectively measure students' abilities.

However, not all test items meet the criteria of a good assessment instrument. Some may be too easy or too difficult, fail to distinguish between high- and low-achieving students, or contain distractors that are ineffective. Therefore, conducting item analysis is important to assess the quality of the questions, especially in terms of difficulty level, discrimination index, and distractor effectiveness.

At SMP Negeri 4 Pematangsiantar, teachers are responsible for ensuring that the test instruments they use effectively reflect students' competencies. This observation aims to evaluate the quality of the test items used in Class 7A by analyzing them based on established evaluation principles. The results of this study are expected to provide meaningful insights and recommendations for improving future test designs.

RESEARCH METHOD

Research Design

This study employed a quantitative research design, focusing on the analysis of test items administered to Class 7A students at SMP Negeri 4 Pematangsiantar. Quantitative methods were chosen to enable the collection and analysis of numerical data, providing an objective evaluation of the test instruments' quality in terms of validity, reliability, practicality, authenticity, and washback.

Research Participants

The participants in this study were 14 students from Class 7A of SMP Negeri 4 Pematangsiantar during the academic year 2024/2025. This class was selected due to its relevance to the English language learning materials and the accessibility for observation and data collection.

Research Instruments

The primary instrument used in this research was a set of English language proficiency tests covering four major language skills: listening, speaking, reading, and writing. Each skill was assessed through 30 multiple-choice items and 15 essay questions, resulting in a total

of 180 questions. The test items were adapted from the textbook English for Nusantara Grade 7, ensuring alignment with the curriculum and students' cognitive levels.

The tests were developed to meet assessment principles:

- Validity, by ensuring alignment between test items and learning objectives.
- Reliability, through careful item construction and standardized scoring rubrics.
- Practicality, by designing tests feasible to administer under real classroom conditions.

Data Collection Procedures

Data were collected during the Daily Assessment (Penilaian Harian) in English class. The researcher observed the test administration directly without intervening, using a structured observation sheet to record students' behaviors, level of concentration, and difficulties encountered during the test. Students' answers were then collected for further analysis.

Data Analysis Techniques

The collected data were analyzed using the principles of language assessment as proposed by Brown (2004):

1. Validity: Examined through content and construct analysis
2. Reliability: Measured using Cronbach's Alpha with SPSS version 24 to determine internal consistency.
3. Practicality: Observed through time allocation, cost efficiency, and resource availability.
4. Authenticity: Evaluated by the degree to which test items reflected real-world language use.
5. Washback: Assessed by identifying how the tests influenced student learning and motivation.

Both quantitative analysis (descriptive statistics) and qualitative observations were used to provide a comprehensive evaluation of the test instruments.

RESULT AND DISCUSSION

Findings Based on Practicality

The practicality of the test was evaluated in terms of administration time, availability of resources, and cost efficiency. Based on the observation and documentation during the test implementation, the test was considered fairly practical, although it required several logistical and technical considerations. The test was conducted over two consecutive days,

with two hours allocated each day, to accommodate all four language skills (Listening, Speaking, Reading, and Writing). This schedule allowed students to complete the tasks with better focus and reduced fatigue, ensuring optimal performance for each skill. For the Listening section, the researcher provided a personal speaker to ensure the clarity and quality of the audio recordings used in the test. This was necessary due to the school's limited audio equipment. While this required additional preparation from the researcher, the use of personal equipment allowed the test to proceed smoothly.

Overall, the administration of the test demonstrated a reasonable level of practicality, despite requiring external technical support and flexible scheduling. The test could be conducted effectively and efficiently under real classroom conditions without causing significant logistical or financial burden.

Finding Based on Reliability

1. Reliability Writing Skill

   Reliability for the multiple-choice section of the writing skill test was high, with a Cronbach's Alpha of 0.857, indicating strong internal consistency. This suggests that the writing multiple-choice items consistently measure what they intend to measure across different students.

   However, for the essay section, the reliability was lower, with a Cronbach's Alpha of 0.688, which is slightly below the standard threshold of 0.70. This indicates that some items may need refinement or clearer rubrics to improve consistency in scoring.

2. Reliability Reading Skill

   The multiple-choice section of the reading skill showed very high reliability with a Cronbach's Alpha of 0.875, suggesting that the test is highly consistent and dependable. Meanwhile, the essay section had a Cronbach's Alpha of 0.650, showing moderate reliability. Although acceptable, the results indicate a need to enhance scoring consistency, potentially through improved training or more detailed rubrics for evaluators.

3. Reliability Speaking Skill

   The multiple-choice items for speaking achieved excellent reliability, with a Cronbach's Alpha of 0.900, which strongly supports the consistency of the test items. On the other hand, the essay section scored 0.585, indicating low reliability. This suggests a high variability in how student responses were scored, possibly due to the subjective nature of oral assessments. Improvements in scoring rubrics or rater training are recommended.

4. Reliability Listening Skill

For the multiple-choice listening items, the Cronbach's Alpha was 0.865, demonstrating high reliability. The items were consistently understood and responded to by students.

Interestingly, the essay section showed the highest reliability among all essay tests, with a Cronbach's Alpha of 0.887, indicating that the listening essay items were clearly structured and that scoring was consistent among raters.

Specifically, in the essay section for speaking skills, reliability was found to be the lowest compared to the other language skills. This is most likely due to the subjective nature of oral assessment, which involves aspects such as intonation, pronunciation, fluency, and sentence structure—elements that may be interpreted differently by each rater.

To improve the reliability of future tests, several strategies can be implemented:

1. Develop and apply a more detailed analytic scoring rubric, which includes clearly defined indicators such as fluency, pronunciation, grammatical accuracy, and coherence.
2. Conduct training for raters to increase consistency and develop a shared understanding of the scoring criteria (inter-rater reliability).
3. Involve more than one rater in the scoring process for speaking essays and calculate inter-rater reliability (e.g., using Pearson correlation) to ensure scoring consistency across raters.
4. Provide benchmark samples—standardized student responses with assigned scores—to serve as scoring references, thereby promoting more objective and consistent evaluations.

In conclusion, while the multiple-choice sections of the test demonstrated very high reliability, the essay sections, particularly in speaking, still require improvement in test design and scoring methods. The improvement strategies outlined above are expected to support a more consistent, accurate, and trustworthy assessment process in future implementations.

Table 1. Finding based on reliability

Findings Based on Reliability			
Skill	Test Type	Cronbach's Alpha	Interpretation
Writing	Multiple Choice	0.857	High reliability
Writing	Essay	0.688	Moderate reliability
Reading	Multiple Choice	0.875	Very high reliability
Reading	Essay	0.650	Moderate reliability
Speaking	Multiple Choice	0.900	Excellent reliability
Speaking	Essay	0.585	Low reliability (needs review)
SBC >0,7= realiability low			

Finding Based on Validity

1. Validity Writing Skill

The results of the validity analysis for the 30 writing skill test items are presented in the table above. Based on the calculations, the r-table value at a 5% significance level with the given number of respondents is 0.552. A test item is considered valid if the r-count is greater than the r-table (r-count > r-table), and invalid if r-count < r-table. These findings indicate that 30% of the items (10 out of 30) are valid, while 70% (20 out of 30) are invalid. As such, further analysis is necessary to revise the invalid items and improve the overall quality of the test instrument.

Based on the validity analysis, only 40% (6 out of 15) of the writing essay items met the required validity criteria. It is therefore recommended that the invalid items be revised or replaced with better-constructed items that reflect the learning indicators and accurately measure students' writing skills. This will help improve the overall quality and accuracy of the test instrument.

2. Validity Reading Skill

The validity test results for the 30 reading skill test items are presented in the table above. Based on the analysis, the r-table value at the 5% significance level is 0.552. An item is considered valid if the r-count is greater than the r-table (r-count > r-table), and invalid if r-count < r-table. Thus, the proportion of valid items is 53.3% (16 items), while the proportion of invalid items is 46.7% (14 items). These results indicate that more than half of the items meet the validity criteria, suggesting moderate overall test quality with room for improvement

Only 40% (6 out of 15) of the reading essay items fulfilled the validity requirements. This suggests that the overall validity level is moderate for the reading section. Therefore, it

is recommended that the invalid items be revised, removed, or rewritten to better reflect the reading objectives and improve the quality and accuracy of the assessment instrument.

3. Validity Speaking Skill

The validity test results for the 30 speaking skill test items are shown in the table above. Based on the analysis, the r-table value at the 5% significance level is 0.532. An item is considered valid if the r-count is greater than the r-table (r-count > r-table), and invalid if r-count < r-table. These results indicate that more than half of the speaking skill items meet the validity standard, making them appropriate for use in assessing students' speaking ability. However, the invalid items require revision or replacement to ensure that the test instrument is of high quality overall.

Only 33% (5 out of 15) of the speaking essay items met the validity criteria. This indicates a low overall validity for the speaking section. To improve test quality, it is strongly recommended that the invalid items be revised or replaced, with greater attention to alignment with speaking performance indicators, such as clarity, fluency, coherence, and pronunciation accuracy. Additionally, clearer task instructions and rubrics may help ensure the test measures the intended speaking skills more effectively.

4. Validity Listening Skill

The validity test results for the 30 listening skill test items are shown in the table above. Based on the analysis, the r-table value at the 5% significance level is 0.532. A test item is considered valid if the r-count is greater than the r-table (r-count > r-table), and invalid if r-count < r-table. These results indicate that more than half of the listening skill items did not meet the validity standard. Therefore, further analysis is needed for the invalid items in order to improve the overall quality of the test instrument.

The results indicate that 80% (12 out of 15) of the essay items in the listening section are valid, which reflects a generally strong alignment between the test items and the intended listening skill constructs. However, it is recommended that the invalid items be carefully reviewed and possibly revised or replaced to enhance the overall quality and construct validity of the test

Finding Based on Authenticity

The authenticity aspect of language assessment refers to the degree to which test items reflect meaningful and real-life language use. Based on the analysis and observation during the test administration, it can be concluded that the test items in this study were

designed with careful consideration of the principle of authenticity as proposed by Brown (2004), which emphasizes communicative tasks that are relevant to learners' real-life language experiences.

In the Listening section, for example, students listened to recorded conversations, short announcements, and simple monologues that mirror authentic situations found in everyday life, particularly in school or social environments. Students were asked to comprehend meaning, identify key information, and make inferences from what they heard.

In the Speaking section, students were given open-ended prompts such as "Describe your best friend" or "Talk about your favorite holiday destination." These types of tasks encouraged students to use the language actively and personally, aligning with real-world oral communication.

The Reading section included short texts such as emails, public notices, or brief informational passages, followed by comprehension questions involving main ideas, supporting details, and vocabulary interpretation. Meanwhile, in the Writing section, students were asked to compose short texts, such as writing an email to a teacher or describing their favorite food, as well as revising sentences into more natural and grammatically correct paragraphs.

The findings indicate that the test items effectively represented authentic language contexts, where students were required to apply their language skills meaningfully, rather than rely solely on rote memorization or isolated grammar rules.

Therefore, it can be concluded that the test instrument fulfilled the criteria of authenticity, as it provided students with real-life communicative tasks, helping to enhance the validity of the test and support the development of functional language skills in a practical learning environment.

Finding Based on Washback

Washback refers to the influence that a test exerts on the teaching and learning process. In this study, the washback effect was examined to determine how the administration of the test impacted students' motivation, learning strategies, and their overall perception of English language learning.

Based on observations and students' responses during and after the test, it was found that the test generated a positive washback effect. Many students demonstrated enthusiasm when faced with challenging tasks, particularly in the Speaking and Writing sections, where they had more freedom to express their thoughts. The test also encouraged students to

prepare more thoroughly—not only through memorization but also by practicing contextual language use, which reflects deeper learning.

Teachers also benefited from the test results, which provided valuable insights into students' strengths and weaknesses across the four language skills. These insights enabled teachers to evaluate the effectiveness of their teaching methods and adjust instructional strategies, materials, and classroom activities accordingly.

However, some negative washback effects were also noted. A few students experienced anxiety, especially during the Speaking and Listening tasks, which required higher levels of concentration and confidence. Time limitations also created additional pressure. Nevertheless, these negative effects were relatively minor and can be minimized through better preparation and more supportive classroom environments.

In general, the test demonstrated a predominantly positive washback, contributing meaningfully to both student learning and instructional development. With continued refinement and responsive adjustments based on feedback, the test can serve as an effective tool for improving English language teaching and assessment in the classroom.

CONCLUSION

This study aimed to evaluate the effectiveness of English language test items used to measure students' skills in Listening, Speaking, Reading, and Writing, based on the principles of language assessment. The test instrument, consisting of 120 multiple-choice items and 60 essay questions, was designed and analyzed in terms of validity, reliability, practicality, authenticity, and washback.

The findings revealed that only a portion of the test items met the criteria for validity, with writing and listening skills showing lower proportions of valid items, indicating a need for item revision. Reliability analysis confirmed that the instrument produced consistent results over time, especially in the multiple-choice section. The practicality of the test was supported by manageable logistics, including time allocation, cost, and resource availability, despite minor technical needs such as the use of speakers for the listening section.

In terms of authenticity, the test reflected real-life language use, particularly in the speaking and writing tasks that required students to communicate ideas relevant to daily situations. The test also generated positive washback, motivating students to engage more seriously with the learning material, while giving teachers meaningful feedback for instructional improvement.

Overall, the test instrument provided valuable insights into students' English proficiency and demonstrated a balanced application of language assessment principles.

However, further refinement—especially in revising invalid items and enhancing speaking and listening assessments—is recommended to increase the overall quality and effectiveness of the evaluation process.

REFERENCES

Abdurrasyid, A., & Panggabean, H. S. (2024). Steps in implementing the evaluation of Islamic Religious Education (PAI) learning. *JIM: Jurnal Ilmiah Mahasiswa Pendidikan Sejarah*, *9*(4), 835-841.

Aeni, E. S., Wuryani, W., & Rostikawati, Y. (2019). Penerapan metode Copy The Master pada pembelajaran menulis teks argumentasi untuk meningkatkan kreativitas menulis mahasiswa. *Diglosia–Jurnal Pendidikan, Kebahasaan, dan Kesusastraan Indonesia*, *3*(2), 50-65.

Akidah, I., & Mansyur, U. (2022). Strategi image streaming terhadap kemampuan menulis pada mahasiswa. *Literasi: Jurnal Bahasa dan Sastra Indonesia serta Pembelajarannya*, *6*(2), 406-413.

Amalia, R. (2018). Penerapan Strategi Copy The Master Dalam Meningkatkan Keterampilan Menulis Puisi Pada Siswa Kelas VIII. A Smp Negeri 3 Labakkang Kabupaten Pangkep. *Skripsi. Universitas Muhammadiyah Makassar, Hal-22*.

Fitriani, N., & Rahmadewi, S. (2025). Development and Role of Measurement Tools in Educational Evaluation. *JMPI: Jurnal Manajemen, Pendidikan dan Pemikiran Islam*, *3*(1), 78-97.

Hayati, U., Ediyani, M., Maimun, M., Anwar, K., Fauzi, M. B., & Suryati, S. (2020). Test technique as a tool for evaluation of learning outcomes. *Budapest International Research and Critics Institute-Journal (BIRCI-Journal)*, *3*(2), 1198-1205.

Marganingsih, M. (2022). Peningkatan Keterampilan Menulis Cerpen Melalui Media Teks Lagu Dengan Metode Latihan Terbimbing. *Jurnal Ilmiah Bahasa dan Sastra*, *6*(6).

Ningrum, M., & Rabiah, S. (2022). Model Konstruktivistik Terhadap Peningkatan Kemampuan Menulis Teks Ceramah Siswa. *Journal of Language and Literature*, *2*(2), 180-187.

Nurhayati. 2019. Apresiasi Prosa Fiksi Indonesia (Revisi). Jawa Tengah: Cakrawala Media.

Nurjannah, A., & Suhara, A. M. (2019). Analisis penggunaan bahasa daerah dalam pembelajaran menulis cerpen di kelas ix smpn 1 cipatat kabupaten bandung barat. *Parole: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, *2*(2), 255-262.

Pertiwi, S., & Kolen, K. V. (2020). Pengaruh Media Film Terhadap Keterampilan Menulis Narasi Pada Mata pelajaran Bahasa Indonesia Pada Siswa Kelas V SD 02 Pagi Cipayung. *Jurnal Inovasi Pendidikan MH Thamrin*, *4*(1), 10-19. Pertiwi, S., & Kolen, K. V. (2020). Pengaruh

Media Film Terhadap Keterampilan Menulis Narasi Pada Mata pelajaran Bahasa Indonesia Pada Siswa Kelas V SD 02 Pagi Cipayung. *Jurnal Inovasi Pendidikan MH Thamrin*, *4*(1), 10-19.

Prasad, M. (2024). Introduction to the GRADE tool for rating certainty in evidence and recommendations. *Clinical Epidemiology and Global Health*, *25*, 101484.

Puspita, S. M., & Syihabuddin, S. (2020). Forms of Instruments in Assesing Vocabulary Mastery. *International Journal of Arabic Language Teaching*, *2*(02), 213-231.

Retno, R. S., Purnomo, P., Hidayat, A., & Mashfufah, A. (2025). Conceptual framework design for STEM-integrated project-based learning (PjBL-STEM) for elementary schools. *Asian Education and Development Studies*, *14*(3), 579-604.

Wambugu, P. W., & Changeiywo, J. M. (2008). Effects of mastery learning approach on secondary school students' physics achievement. *Eurasia Journal of Mathematics, Science and Technology Education*, *4*(3), 293-302.