INNOVATIVE: Journal Of Social Science Research
Volume 5 Nomor 4 Tahun 2025 Page 9604-9619
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Implementation of English Test Based on The Principles of Language Assessment and Evaluation at SMP Negeri 2 Bandar Perdagangan

Leony Elisabeth Situmorang[1], Riyanti Anjelina Sinaga[2], Naomi Arta Tambunan[3], Adika Indotua Nainggolan[4], Dumaris E. Silalahi[5]✉
Universitas HKBP Nommensen Pematangsiantar
Email: dumaris.silalahi@uhnp.ac.id[5]✉

Abstrak

Penelitian ini mengkaji implementasi ujian bahasa Inggris berdasarkan lima prinsip penilaian bahasa: kepraktisan, keandalan, validitas, keaslian, dan efek balik. Ujian tersebut mencakup keterampilan mendengarkan, berbicara, membaca, dan menulis melalui soal pilihan ganda dan tugas esai yang diberikan kepada 26 siswa kelas tujuh di SMP Negeri 2 Bandar Perdagangan. Desain kualitatif deskriptif disertai data kuantitatif digunakan, termasuk pemeriksaan keandalan antarpenilai dan analisis kesulitan soal. Hasil penelitian menunjukkan bahwa analisis menemukan ujian ini praktis untuk kondisi kelas nyata, andal melalui penilaian ganda yang konsisten, valid dalam mengukur keterampilan yang dimaksud dengan tugas-tugas autentik yang mencerminkan konteks kehidupan nyata, dan menciptakan dampak positif dengan memberikan umpan balik yang berguna untuk perbaikan pengajaran dan pembelajaran. Selain itu, temuan ini menjelaskan peran akurasi gramatikal dalam kinerja siswa, terutama untuk keterampilan produktif. Studi ini menunjukkan bahwa prinsip-prinsip evaluasi penilaian bahasa dapat diterapkan secara efektif di kelas Bahasa Indonesia sebagai Bahasa Asing (EFL), yang memastikan penilaian yang bermakna dan adil yang mendukung perkembangan bahasa siswa.

Kata Kunci: *Pelaksanaan, Ujian Bahasa Inggris, Prinsip-Prinsip, Penilaian dan Evaluasi Bahasa*

Abstract

This study investigates the implementation of an English test based on the five principles of language assessment: practicality, reliability, validity, authenticity, and washback. The test covered listening, speaking, reading, and writing skills through multiple-choice and essay tasks which administered to 26 students of the seventh grade students at SMP Negeri 2 Bandar Perdagangan. A descriptive qualitative design supporting by quantitative deta support is used, involving inter-rater reliability checks and item difficulty analysis. Findings show that the analysis finds the practical test for real classroom conditions, reliable through consistent double scoring, valid in measuring intended skills with authentic tasks reflecting real-life contexts, and created positive washback by providing useful feedback for teaching and learning improvement. Otherwise the finding elaborates the role of grammatical accuracy in students' performance, especially for productive skills. This study indicates that language assessment evaluation principles can be applied effectively in Indonesian as EFL classrooms, which ensuring meaningful and fairly testing that supports students' language development.

Keywords: *Implementation, English Test, Principles, Language Assessment and Evaluation*

INTRODUCTION

English has become one of the compulsory subjects taught in Indonesian junior high schools, playing an important role in developing students' ability to communicate effectively in both spoken and written forms (Brown, 2004). The National Curriculum demands that students achieve basic competencies in listening, speaking, reading, and writing to help them participate in the global community. However, despite this objective, the reality in many schools, including SMP Negeri 2 Bandar Perdagangan, shows that English language learning and testing often do not fully measure students' actual skills. Teachers frequently use test items that focus on vocabulary memorization and grammar rules in isolation, while neglecting the integration of skills in real-life contexts.

One of the most common challenges faced by English teachers in Indonesian junior high schools is the design and implementation of language tests that meet international standards of assessment. Teachers sometimes have limited understanding of test construction principles due to lack of training and limited access to updated resources (Hughes, 2003). In rural schools, these challenges are even more evident because of restricted facilities such as the absence of language laboratories, limited internet access, and time constraints during teaching hours. As a result, the tests tend to emphasize rote learning rather than authentic use of language, which may lead to inaccurate evaluation of students' real abilities.

The importance of effective language assessment in English as a Foreign Language (EFL) classrooms has been widely emphasized in recent literature. Brown (2004), Hughes (2003), and Fulcher (2010) have outlined five fundamental principles practicality, reliability, validity, authenticity, and washback that should guide the construction and implementation of language tests. These principles ensure that language assessments are not only fair and consistent, but also reflect real-life language use and promote positive effects on teaching and learning practices. However, in many Indonesian junior high schools, especially those located in rural areas, English assessments often fail to reflect these principles. As revealed by Situmorang et al. (2025), test practices at SMP Negeri 2 Bandar Perdagangan frequently focused on isolated grammar and vocabulary, neglecting integrated language skills and real-world communication. This results in limited student engagement and inaccurate representations of learners' actual language competence.

Supporting this concern, Sinaga, Sagala, and Silalahi (2025) explored the implementation of sustainable assessment practices in another junior high school context and found that traditional test formats particularly those limited to written tests were inadequate in developing comprehensive English proficiency. Their study highlighted students' desire for more varied, authentic, and engaging assessment methods such as oral presentations, group discussions, and task-based evaluations. Furthermore, they emphasized the importance of timely and constructive feedback in motivating students and fostering self-reflection. These findings align with the washback principle proposed by Brown (2004), which advocates for assessments that inform and enhance learning rather than merely measure it.

Both studies underscore the need for Indonesian EFL teachers to go beyond conventional testing models by developing classroom-based assessments that are both theoretically grounded and practically feasible. They also call for increased teacher training and resource support to bridge the gap between assessment theory and classroom application. By implementing assessments rooted in core language assessment principles and supported by sustainable practices, educators can foster more meaningful and accurate evaluations of student learning, while simultaneously improving English proficiency and classroom engagement.

Language assessment is an integral part of the language teaching process. Brown (2004) highlights that testing is not only an instrument for measuring students' achievement but also an essential tool to guide teachers in making instructional decisions. Hughes (2003) defines language testing as the process of measuring a learner's ability in language using

various techniques that are fair, practical, reliable, and valid. Fulcher (2010) explains that the design of a language test should be based on five main principles: practicality, reliability, validity, authenticity, and washback. Practicality refers to the feasibility of administering the test under real classroom conditions, while reliability concerns the consistency and dependability of the scores obtained. Validity means the test actually measures what it is supposed to measure. Authenticity involves the use of real-life contexts that reflect actual language use. Finally, washback is the impact that the test has on teaching and learning practices.

Several previous studies have explored these principles and emphasized their importance in effective language assessment. Siahaan (2023) found that implementing practical and authentic test tasks increases student motivation in EFL classrooms. Similarly, Fulcher (2010) and Hughes (2003) provided various theoretical perspectives on how teachers can design tests that are aligned with curriculum goals and communicative needs. However, despite these insights, there is still limited empirical research focusing on the actual implementation of tests that simultaneously measure the four English skills in Indonesian junior high schools, especially in rural areas with limited teaching resources. Most existing research tends to focus on either test theory or the evaluation of ready-made standardized tests rather than the development and trial of new test instruments designed by teachers themselves.

The gap in the practical application of language assessment principles in real classroom settings highlights the need for more context-specific research. This study responds to that need by developing and implementing an English test based on the five fundamental principles of language assessment, focusing on seventh-grade students at SMP Negeri 2 Bandar Perdagangan. The test includes both multiple-choice and essay items covering listening, speaking, reading, and writing skills. It also applies inter-rater reliability checks to ensure fair and objective scoring. This is particularly important since students are often more comfortable answering multiple-choice questions but may struggle with essay tasks, which require them to produce grammatically accurate sentences. By emphasizing grammatical accuracy, especially in productive skills, this test aims to provide a more comprehensive measurement of students' true language competence.

The main objective of this research is to examine how the principles of practicality, reliability, validity, authenticity, and washback can be integrated into a single test format that is feasible for use in junior high school English classrooms. This study is expected to contribute practical insights for teachers who need adaptable and realistic models of English

tests that align with national curriculum standards while remaining practical under school constraints. Moreover, the findings will help future researchers develop similar test instruments that can be applied in different contexts.

In line with this objective, this paper presents the background and theoretical foundations for the test design, describes the research method used to implement and analyze the test, discusses the findings related to each principle of language assessment, and concludes with recommendations for improving the quality of English testing in Indonesian junior high schools. It is hoped that this research will encourage teachers to design their own tests that are valid, reliable, authentic, and capable of producing positive washback, ultimately enhancing students' English proficiency in a meaningful way.

RESEARCH METHOD

This research applied a descriptive qualitative approach supported by quantitative data to provide a clear and in-depth understanding of how an English test based on the principles of language assessment was developed and implemented. The study was conducted at SMP Negeri 2 Bandar Perdagangan with a total of 26 students from class VII-4 as the research participants. These students were selected because they represented typical junior high school learners who face challenges in mastering English language skills.

The instrument used in this study was an English language test specifically designed to assess students' listening, speaking, reading, and writing skills. The test consisted of 30 multiple-choice items and 15 essay questions for each skill. All test items were developed based on the five principles of language assessment, namely practicality, reliability, validity, authenticity, and washback, as formulated by Brown (2004), Hughes (2003), and Fulcher (2010). The test items were also constructed referring to the Indonesian 2013 Curriculum (Kurikulum 2013) to ensure relevance and alignment with national standards. The items covered topics related to students' daily lives, ensuring authenticity, while the structure and instructions were adapted to match the students' level to maintain practicality.

The test was administered during regular English class hours in the students' classroom under the supervision of the English teacher and the researcher. Permission for test administration was granted by the school principal, and all students participated voluntarily. Data were collected through the students' test responses, which were then analyzed using both descriptive and statistical methods. The practicality of the test was evaluated based on its feasibility under existing school conditions and the students' level of understanding. Reliability was examined through the parallel form approach, in which two raters the

researcher and a colleague independently scored the students' answers and compared the results for consistency. The item difficulty index was also calculated to ensure the appropriateness of each test item. Validity was assessed through item analysis using the Pearson Product-Moment Correlation Coefficient (R-count) to measure the correlation between each item score and the total test score, following the guidelines by Karl Pearson (1896). Authenticity was analyzed by reviewing the test content to ensure it reflected real-life contexts that students might encounter. Washback was analyzed by observing the test's impact on students' motivation and the teacher's reflection for future instructional improvement.

In scoring the test, multiple-choice items were given a score of one point for each correct answer. Essay questions for writing and speaking were scored based on the accuracy of students' responses, with a maximum score of two for correct answers and one for partially correct responses to acknowledge students' effort and encourage participation. For listening and reading essay tasks, answers written in English were given full credit when accurate. This scoring approach highlighted grammatical accuracy as an essential part of measuring students' productive language skills.

The results obtained were tabulated and interpreted in line with the five assessment principles to draw conclusions about the effectiveness of the test design. The data served not only to measure students' performance but also to provide insights into how well the test aligned with the principles of practicality, reliability, validity, authenticity, and washback, and how it could inform better English language assessment practices for junior high school students.

RESULT AND DISCUSSION

Student Test Results

The English test was administered to 26 seventh-grade students at SMP Negeri 2 Bandar Perdagangan to measure their listening, speaking, reading, and writing skills. The test consisted of 30 multiple-choice questions and 15 essay questions for each skill. The students completed the test within regular lesson hours under teacher supervision.

The results show that students performed better in receptive skills (listening and reading) than in productive skills (speaking and writing). The summary below presents the average, highest, and lowest scores for each skill.

Table 1.

SKILL	Average Score	Highest Score	Lowest Score
Listening	81.1	98	52
Speaking	46.4	57	27
Reading	54.3	100	7
Writing	47.5	85	26

The table shows that listening has the highest average score (81.1), while speaking and writing have lower averages (46.4 and 47.5, respectively). This reflects the common challenge among junior high school students in using English productively, mainly due to limited vocabulary and grammatical accuracy. Reading showed varied results, with the highest score reaching 100 but with a significant gap down to the lowest score of 7, indicating differences in students' reading comprehension levels.

This overview demonstrates the real language ability of the students and highlights the need for reliable, valid, and authentic assessments that address both receptive and productive skills.

PRACTICALITY

The practicality of the test was demonstrated through its easy implementation in a real junior high school classroom with limited facilities. The test was printed using standard A4 paper and did not require any expensive equipment or a language laboratory. For the listening section, the teacher read the listening script aloud or used a simple speaker to play audio, making the test feasible for schools that do not have specialized audio tools.

The entire test, which covered listening, speaking, reading, and writing skills, was conducted within the normal English lesson hours. On average, students completed the test in 90 to 120 minutes. This made the test practical because it did not disturb other class schedules and could be repeated in similar conditions in other classes.

The multiple-choice items were easy to check with an answer key, saving teachers time and effort in scoring. For essay questions, clear scoring rubrics helped teachers mark answers fairly and quickly. This reduced the burden on teachers, ensuring that the test could be used regularly without adding excessive workload.

From the students' perspective, the instructions for each section were adapted to their English level. The tasks used familiar contexts, such as school announcements, personal messages, and daily conversations, which made the questions understandable and relevant.

Although some students found writing and speaking tasks more challenging, they were still able to complete them because the tasks were connected to their real experiences.

In conclusion, the test was practical because it required minimal resources, could be administered within regular class time, and was feasible for both teachers and students.

RELIABILITY

Reliability is a critical aspect of language testing because it shows whether the test can produce consistent results over time and across different raters (Brown, 2004; Hughes, 2003). This study established reliability through inter-rater reliability for productive skills (Writing and Speaking) and split-half reliability with the Spearman-Brown prophecy formula for all skills tested.

Inter-rater Reliability

To check scoring consistency for subjective tasks, the researcher and a peer rater independently scored the Essay Writing and Essay Speaking sections using the same rubric.

This table presents a detailed comparison of the scores.

Table 2.

No	Student Name	ESSAY WRITING		GAP	ESSAY SPEAKING		GAP
		LEONY	ADIKA		LEONY	ADIKA	
1	Niko Al-Aufpar	22	23	1	23	24	1
2	Paul Sitompul	22	22	0	22	22	0
3	Alfredo Tambunan	28	26	2	28	28	0
4	Atha Rizky	26	26	0	21	24	3
5	Arya Sinulingga	28	28	0	22	22	0
6	Nazwa Zafira Purba	28	29	1	28	29	1
7	Wilson Sitorus	28	29	1	24	25	1
8	Chika Tasya Aritonang	27	27	0	25	25	0
9	Gabriel Kresna Situmorang	25	26	1	30	31	1
10	Irwan Simbolon	25	25	0	21	21	0
11	Ibnu Fandra	25	25	0	24	25	1
12	Keyla Marsonna	24	25	1	23	24	1
13	Alya Syahfitri	27	28	1	25	26	1
14	Kollose Nainggolan	20	21	1	18	18	0
15	Bramudya Al Fasyah	24	25	1	22	23	1
16	Syahwana	24	24	0	20	20	0

17	Paskah Tetty N. Nainggolan	26	27	1	27	28	1
18	Jesikha Sidabutar	27	28	1	28	29	1
19	Rafa Andre A. Sinaga	27	28	1	25	26	1
20	Klara S	30	31	1	26	27	1
21	Yolanda Napitupulu	28	29	1	29	30	1
22	Rahel Sinaga	28	28	0	21	22	1
23	Rio Siburian	26	27	1	28	29	1
24	Angel Rina Sijabat	29	30	1	26	27	1
25	Ostin Yohana	55	56	1	25	26	1
26	Darwis A. Sorgana	24	24	0	22	23	1

The score gaps between the two raters ranged from 0 to 3 points, with most students showing only 0 or 1 point difference. According to Brown (2004) and Cohen, Manion, & Morrison (2018), such small differences prove that both raters applied the rubric consistently and fairly, with minimal subjectivity. This demonstrates strong inter-rater reliability, meaning that the scoring process was practical, objective, and credible for classroom use.

Split-half and Spearman-Brown Reliability

In addition to inter-rater reliability, the study also calculated split-half reliability to measure internal consistency within each test section. The split-half method used the Pearson Product-Moment correlation, then the scores were adjusted using the Spearman-Brown prophecy formula to account for the full test length (Fulcher & Davidson, 2007).

The results for each skill were as follows:

- Listening: Pearson r = 0.7059, Spearman-Brown = 0.8277
- Reading: Pearson r = 0.8088, Spearman-Brown = 0.8943
- Writing: Pearson r = 0.8011, Spearman-Brown = 0.8896
- Speaking: Pearson r = 0.8327, Spearman-Brown = 0.9087

All coefficients exceed the widely accepted minimum reliability threshold of 0.70, showing that each section consistently measures what it is intended to measure. The high Spearman-Brown values confirm that if the test were repeated under similar conditions, the results would remain stable.

These findings confirm that the test is both internally consistent and objectively scored, which strengthens the credibility and fairness of the overall assessment. Therefore, the reliability measures demonstrate that the test can be confidently used for evaluating students' English proficiency in real classroom settings.

VALIDITY

A test must demonstrate validity to ensure that it accurately measures the skills it intends to assess. In this study, the validity of the English test was examined through three types of evidence: content validity, construct validity, and empirical validity.

Content Validity

Content validity refers to the degree to which the test content represents the targeted language skills and matches the curriculum requirements. As recommended by Brown (2004) and aligned with the 2013 Curriculum (Permendikbud No. 37 of 2018), the essay writing and essay speaking tasks were designed to reflect the Kompetensi Dasar (Basic Competencies) for junior high school students.

For example, in Grade VII, *KD 4.12* requires students to write simple descriptive texts about people, animals, or objects in their surroundings using appropriate vocabulary and grammar. To meet this, each essay prompt in the test asked students to write a short paragraph about familiar topics such as daily routines, family members, or favorite activities. The prompts were reviewed and judged by a peer teacher and a supervising lecturer to ensure that they aligned with the curriculum objectives, fulfilling the principle of logical validation (Cohen, Manion, & Morrison, 2018).

This expert judgment process guaranteed that the essay tasks truly measured the students' ability to produce descriptive texts as required, ensuring that content validity was strong and practically applicable in classroom contexts.

Construct Validity

Construct validity examines whether the test items properly measure the intended construct in this case, the students' listening, speaking, reading, and writing abilities. This was analyzed statistically using the Pearson Product-Moment Correlation Coefficient, following the method proposed by Karl Pearson (1896) and explained by Brown (2004). The correlation coefficient (*R-Count*) for each item was compared to the R-Table value for $n = 26$ and $\alpha = 0.05$, which is 0.388.

Summary of Construct Validity Analysis

- Listening Skill: Out of 30 items, 22 items were valid (*R-Count* > 0.388) and 8 were invalid. For example, *Q16* (0.709), *Q18* (0.651), and *Q28* (0.872) were strongly valid.
- Speaking Skill: Out of 30 items, 27 items were valid and 3 were invalid. For instance, *Q1* (0.853) and *Q22* (0.942) showed strong positive correlations.

- Reading Skill: Out of 30 items, 14 items were valid and 16 were invalid. Valid examples include *Q15* (0.597) and *Q19* (0.843).
- Writing Skill: Out of 30 items, 28 items were valid and only 2 were invalid, with high correlations such as *Q5* (0.929) and *Q27* (0.906).

This statistical evidence confirms that most items successfully measured the intended skills, demonstrating strong construct validity supported by empirical correlation.

Empirical Validity

Empirical validity was strengthened by conducting an item analysis of 10 representative multiple-choice items for each skill. The difficulty index (*P Value*) indicates the proportion of students who answered each item correctly, interpreted using standard criteria (Brown, 2004).

Item Difficulty Index (P) for Listening Section

Question	Number Correct (T)	P Value	Category
Q10	16	0.62	Moderate
Q12	16	0.62	Moderate
Q14	11	0.42	Moderate
Q16	22	0.85	Easy
Q18	23	0.88	Easy
Q20	17	0.65	Moderate
Q22	15	0.58	Moderate
Q24	11	0.42	Moderate
Q26	18	0.69	Moderate

Item Difficulty Index (P) for Speaking Section

Question	Number Correct (T)	P Value	Category
Q3	13	0.50	Moderate
Q6	20	0.77	Easy
Q9	21	0.81	Easy
Q12	21	0.81	Easy
Q14	19	0.73	Easy
Q16	19	0.73	Easy
Q18	19	0.73	Easy
Q20	11	0.42	Moderate
Q22	15	0.58	Moderate

Item Difficulty Index (P) for Reading Section

Question	Number Correct	P Value	Category
Q2	21	0.81	Easy
Q4	16	0.62	Moderate
Q6	16	0.62	Moderate
Q8	19	0.73	Easy
Q10	14	0.54	Moderate
Q12	11	0.42	Moderate
Q14	2	0.08	Difficult
Q16	9	0.35	Moderate
Q18	21	0.81	Easy

Summary of Item Difficulty Index:

- Listening: P-values ranged from 0.42 (Moderate) to 0.88 (Easy). For example, *Q18* (0.88) was classified as easy.
- Speaking: Most items were easy, with P-values up to 0.81 (*Q9*, *Q12*), and a few moderate items such as *Q20* (0.42).
- Reading: Items varied between moderate (e.g., *Q15* = 0.58) and easy (*Q23* = 0.38).
- Writing: The writing MC items also showed a range, with most moderate (*Q4* = 0.62) and some easy (*Q2* = 0.81). Only *Q14* (0.08) was classified as difficult, indicating that this item might be too challenging for the tested students.

The distribution of difficulty levels shows that the test items were neither too easy nor too difficult overall, which supports the test's fairness and practical use in assessing students with varying proficiency levels.

AUTHENTICITY

Authenticity in language assessment means that test tasks reflect real-life language use and communicative situations students may face daily (Brown & Abeywickrama, 2019). In this study, authenticity was embedded purposefully in the reading, listening, writing, and speaking tasks.

The test items used familiar, age-appropriate topics and situations connected to junior high school students' daily lives. For example, reading texts resembled school announcements, class schedules, or short messages; listening passages were simple daily conversations or announcements; writing prompts asked about daily routines or family members; and speaking tasks involved short self-introductions or describing hobbies. This

approach ensured students practiced language in meaningful, relatable contexts rather than isolated grammar drills.

The tasks focused on functional and purposeful language use such as describing, informing, or inviting, rather than only testing forms. Although some students used partial L1 (Bahasa Indonesia) in their answers because of limited English skills, they still demonstrated the intended communicative purpose. For example, *"Halo, nama saya Alfredo. Umur saya 13 tahun,"* shows the student could organize ideas and complete a self-introduction meaningfully.

However, some students struggled to respond authentically due to limited vocabulary, lack of confidence, fatigue from test duration, or classroom distractions. A few left essay questions blank or needed repeated instructions. This indicates that authentic tasks alone do not guarantee successful performance without adequate support and scaffolding.

To keep tasks fair and practical, the prompts were written in simple language, sometimes clarified in both English and L1. The scoring rubric allowed partial credit for mixed-language responses to encourage students to express real information without fear of penalty. Age-appropriate contexts also helped students relate the tasks to their lives.

Overall, the test demonstrated thematic and situational authenticity by mirroring real communicative situations and encouraging students to use English for practical purposes. However, authentic assessment should always be paired with clear instructions, supportive classroom management, and regular practice to build students' readiness and confidence.

WASHBACK

Washback is an essential principle in language assessment that refers to the influence a test has on the teaching and learning process. Positive washback can motivate students, inform teachers, and guide more effective classroom practices, while negative washback may cause stress, disengagement, or unproductive learning behaviors (Brown & Abeywickrama, 2019).

In this study, the washback effect was observed through student reactions, classroom observations during the trial, and the researcher's reflection as the test facilitator. The test was administered to 26 seventh-grade students at SMP Negeri 2 Bandar Perdagangan and assessed all four language skills reading, listening, writing, and speaking using 30 multiple-choice and 15 essay questions per skill. The overall design aimed to measure real language abilities while encouraging both students and teacher to reflect on strengths and weaknesses.

Several positive washback effects emerged. Many students became more aware of which skills needed improvement. Writing and speaking tasks were perceived as more challenging than reading or listening, making students realize where to focus future learning efforts. Some students felt proud when they were able to understand audio clips, answer reading questions correctly, or attempt essay sections even when they needed to mix Bahasa Indonesia and English. Such efforts are valuable first steps toward real communicative competence. For the teacher, the test served as an informal diagnostic tool, providing concrete insights about vocabulary gaps, grammar weaknesses, and students' dependence on L1 when expressing ideas. It also highlighted areas needing more targeted practice, such as listening comprehension and reading fluency. Despite challenges, many students completed all parts of the test with serious effort, showing genuine curiosity about familiar topics like family, daily routines, food, or hobbies.

However, the study also revealed some negative washback effects. A few students lost interest during longer sections, especially the essay parts that required sustained concentration. Many asked for teacher assistance to translate instructions or explain unfamiliar words, showing low confidence in their reading and writing abilities. The large number of items sometimes exceeded the available class time, resulting in unfinished sections and student fatigue. Students with limited vocabulary occasionally felt anxious, which led to incomplete answers or disengagement.

From the facilitator's perspective, the trial offered valuable reflections. The test exposed the students' actual proficiency levels in performing simple communicative tasks, highlighted practical classroom limitations such as time management and student fatigue, and reinforced the need for scaffolding and clear instructions. Flexible scoring rubrics that accepted partial responses in Bahasa Indonesia helped encourage effort rather than penalize students harshly.

Overall, the washback effect was more constructive than harmful. Students gained awareness of their language strengths and weaknesses, and the teacher obtained practical insights for better lesson planning. To strengthen positive washback in future applications, the test should be adjusted to fit realistic class timeframes, instructions should be simplified and clarified bilingually where needed, key vocabulary should be pre-taught, and practice tasks that mirror test scenarios should be integrated into regular lessons. Peer practice for speaking tasks can also help build student confidence.

In summary, this English test did not only assess language skills but also provided meaningful feedback for improving teaching and learning. While some students struggled

with motivation and understanding, many others responded positively and reflected on their capabilities. Therefore, the washback observed in this study was generally beneficial and highlighted the test's potential to support more effective English instruction in junior high school settings.

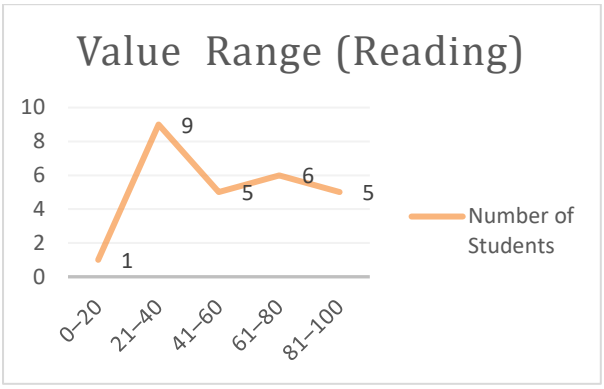

Figure 1. Distribution of Reading Scores

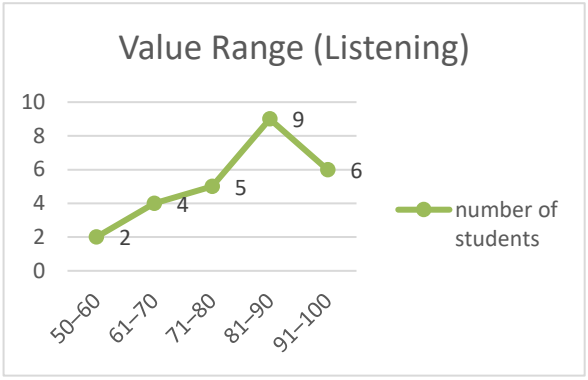

Figure 2. Distribution of Listening Scores

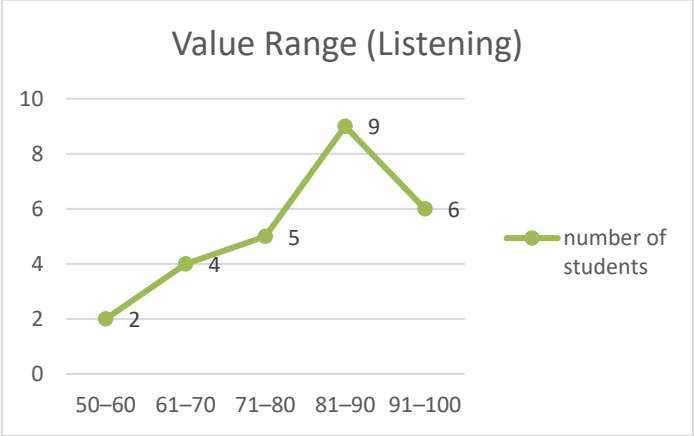

Figure 3. Distribution of Listening Scores

CONCLUSION

This study revealed that the English test developed and implemented at SMP Negeri 2 Bandar Perdagangan successfully applied the five key principles of language assessment: practicality, reliability, validity, authenticity, and washback. The test was found to be practical and feasible for real classroom use, requiring minimal resources and fitting within regular lesson times. Reliability was ensured through consistent scoring by multiple raters and internal consistency analyses, while validity was established through item analysis, expert judgment, and alignment with curriculum standards. The test tasks also demonstrated authenticity by engaging students in meaningful, real-life language use. Moreover, the test created constructive washback by raising students' and teachers' awareness of language proficiency strengths and weaknesses. Although challenges such

as time limitations and student fatigue were observed, the overall assessment design proved to be both effective and informative. This research contributes valuable insights for English teachers and practitioners seeking to develop comprehensive, fair, and context-sensitive assessments that support meaningful learning and teaching improvement in junior high school EFL settings.

REFERENCES

Bitchener, J., & Ferris, D. R. (2015). Written corrective feedback in second language acquisition and writing. Routledge.

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

Brown, H. D., & Abeywickrama, P. (2019). Language assessment: Principles and classroom practices (3rd ed.). Pearson Education.

Creswell, J. W. (2012). Educational research: Planning, conducting, and evaluating quantitative and qualitative research (4th ed.). Pearson.

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

Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2017). Kurikulum 2013: Kompetensi dasar sekolah menengah pertama (SMP). Kemdikbud.

Kirkpatrick, A. (2019). English as an international language in Asia: Implications for language education. Springer.

Purpura, J. E. (2016). Second and foreign language assessment. In G. Hall (Ed.), The Routledge handbook of English language teaching (pp. 374–387). Routledge.

Rahmawati, Y., & Ertin, E. (2020). Developing English speaking assessment instrument for EFL learners. Indonesian Journal of English Education, 7(2), 242–257.

Sinaga, R. S., Sagala, S., & Silalahi, D. E. (2025). Sustainable assessment in improving English proficiency. INNOVATIVE: Journal of Social Science Research, 5(1), 1922–1933.

Sugiyono. (2018). Metode penelitian pendidikan: Pendekatan kuantitatif, kualitatif, dan R&D. Alfabeta.

Tavakoli, P., & Hashemi, M. R. (2019). Assessing speaking and writing: The role of reliability and validity. Language Testing in Asia, 9(1), 1–14.