INNOVATIVE: Journal Of Social Science Research
Volume 5 Nomor 4 Tahun 2025 Page 11195-11203
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

English Test: Assessment And Evaluation Principles Of Grade VII-2 Students At SMP Negeri 2 Siantar

Syalaisha Syafira Irwansyah[1], Venyta Sinaga[2], Viola Sasti Bizora L. Tobing[3], Femy Chandra Winata[4], Dumaris E. Silalahi[5]
Universitas HKBP Nommensen Pematangsiantar
Email: syalaisha05@gmail.com[1]✉

Abstrak

Penelitian ini bertujuan untuk mengevaluasi tes bahasa Inggris berdasarkan lima prinsip penilaian bahasa, yaitu validitas, reliabilitas, kepraktisan, keaslian, dan dampak balikan (washback). Tes diberikan kepada dua puluh empat siswa kelas VII SMP Negeri 2 Pematangsiantar. Konten tes dirancang secara cermat untuk menilai empat keterampilan bahasa Inggris: menyimak, berbicara, membaca, dan menulis. Pendekatan gabungan kualitatif dan kuantitatif digunakan dengan data yang diperoleh melalui rubrik terstruktur, lembar kerja siswa, serta soal pilihan ganda dan esai yang dikembangkan sesuai kerangka Kurikulum Merdeka. Hasil penelitian menunjukkan validitas isi yang tinggi, reliabilitas penilai yang kuat, pelaksanaan yang praktis di kelas, desain tugas yang autentik, serta dampak balikan yang positif. Siswa memperoleh nilai rata-rata lebih tinggi pada keterampilan menyimak (93,5) dan berbicara (90,25), sedangkan membaca (49,5) dan menulis (41,08) menjadi keterampilan yang paling menantang. Temuan ini memberikan wawasan untuk meningkatkan praktik penilaian bahasa di tingkat SMP.

Kata Kunci: *Prinsip penilaian dan evaluasi, Keterampilan bahasa Inggris, Tes bahasa Inggris*

Abstract

This study aims to evaluate an English test based on five language assessment principles: validity, reliability, practicality, authenticity, and washback. The test was administered to twenty-four seventh-grade students at SMP Negeri 2 Pematangsiantar. The content was carefully designed to assess four English skills: listening, speaking, reading, and writing. A blended qualitative and quantitative approach was applied, with data collected through structured rubrics, student worksheets, and multiple-choice/essay tasks developed under the Kurikulum Merdeka framework. The results showed high content validity, strong inter-rater reliability, practical classroom implementation, authentic task design, and positive instructional washback. Students achieved higher average scores in listening (93.5) and speaking (90.25), while reading (49.5) and writing (41.08) were more challenging. These findings provide insights into student performance patterns and support improvements in language assessment practices for junior high school contexts.

Keywords: *Assessment and evaluation principles; English language skills; English test*

INTRODUCTION

Assessment is an essential component of modern language education, particularly in outcome-based learning where academic progress and instructional quality are tied to measurable competencies. In English language teaching (ELT), especially at the junior high school level, effective assessment not only measures achievement but also supports instruction, provides feedback, and fosters communicative competence in line with curriculum standards. Under the Kurikulum Merdeka, English teaching in Indonesia emphasizes the integration of listening, speaking, reading, and writing skills, which requires assessment instruments that are authentic, valid, and practical to ensure accurate evaluation of students' abilities.

Previous studies have outlined core principles of effective assessment, namely validity, reliability, authenticity, practicality, and washback. These principles, discussed collectively by scholars such as Brown, Douglas, and Bachman & Palmer, have been widely applied in standardized tests and specific skill assessments. However, much of the literature focuses on either large-scale testing or isolated language components, leaving a gap in classroom-based research that evaluates both student performance across all four skills and the quality of the test instrument itself. Moreover, existing studies rarely address how such principles can be consistently applied in real teaching contexts within Indonesian junior high schools.

This study seeks to address these limitations by evaluating an English language test administered to Grade VII-2 students at SMP Negeri 2 Siantar, designed to assess four skills using both multiple-choice and essay questions. The novelty lies in its dual focus: measuring student performance and assessing the test instrument against the five assessment principles in a practical classroom setting. The objective of this research is to provide

empirical insights and practical recommendations for improving classroom assessment practices in junior high school English education, ensuring alignment with both theoretical standards and curriculum goals.

RESEARCH METHOD

Research Design

This study applied a descriptive qualitative design supported by quantitative data. The qualitative approach was used to evaluate the test instrument based on the five principles of language assessment—validity, reliability, practicality, authenticity, and washback (Brown & Abeywickrama, 2019). The quantitative approach was used to measure students' performance in four English skills: listening, speaking, reading, and writing.

Research Site and Participants

The research was conducted at SMP Negeri 2 Siantar with a purposive sample of 24 Grade VII-2 students. The participants were selected due to their direct involvement in English instruction and accessibility for testing purposes.

Instruments and Data Collection

The main instrument was an English proficiency test developed by the researcher. The test consisted of four skill-based sections—listening, speaking, reading, and writing—with 30 multiple-choice and 15 essay questions per skill (180 items in total). Listening and reading tasks measured comprehension, while speaking and writing tasks required language production.

Data were collected in two forms:

1. Quantitative data from student scores in all four skills.
2. Qualitative data from an in-depth evaluation of the test design using the five principles of language assessment.

Research Procedure

The test was administered in stages during regular class sessions:

1. Listening tasks were delivered through speakers or teacher read-alouds.
2. Reading and writing tasks were completed on printed sheets.
3. Speaking tasks were assessed individually using structured prompts.

Scoring for essay and speaking tasks used rubric-based guidelines. Two raters (the

researcher and a peer assessor) evaluated the responses independently to ensure inter-rater reliability.

Data Analysis

Quantitative data were analyzed using descriptive statistics to determine the mean, highest, and lowest scores for each skill. Qualitative data were analyzed by reviewing the alignment of the test with curriculum indicators, evaluating scoring consistency, and assessing authenticity, practicality, and washback effects. This combined analysis provided a comprehensive understanding of both student achievement and test quality.

Research Procedure Flowchart

The overall research procedure followed a structured sequence to ensure both validity of data collection and clarity in execution. This procedural framework, which covers the stages from planning to conclusion, is illustrated in Figure 1. The flowchart visually outlines the logical progression of activities, including the research planning, design of the test instrument, participant selection, test administration, scoring, data compilation, data analysis, and final interpretation of findings.

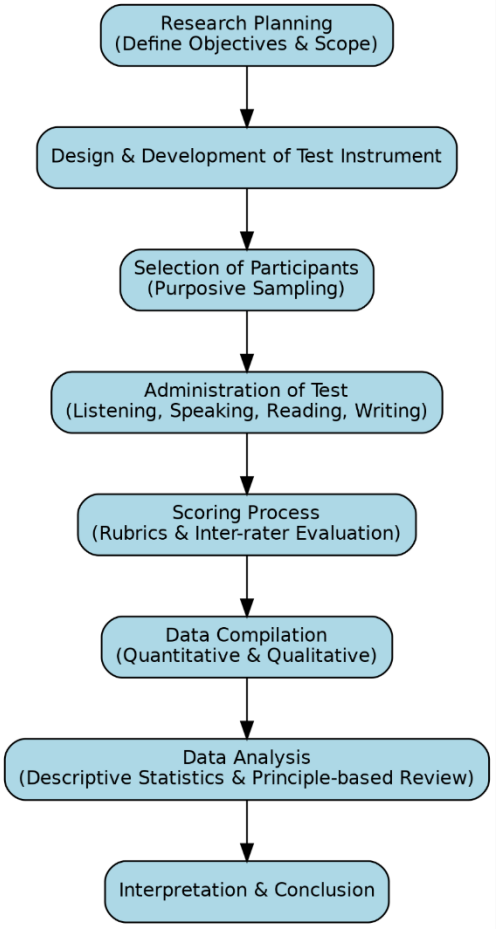

Figure 1. Research Procedure Flowchart

RESULT AND DISCUSSION

Student Performance in English Language Skills

This study assessed the English skills of 24 Grade VII-2 students at SMP Negeri 2 Siantar through a classroom-based test consisting of four components: listening, reading, speaking, and writing. Each skill was assessed using multiple-choice and essay items, supported by a rubric for productive skills. The aim was to evaluate the implementation of five fundamental language assessment principles: validity, reliability, practicality, authenticity, and washback.

The results of student performance are presented in the following table.

No	Name	Listening	Reading	Speaking	Writing	Total
1	Alfahri	92	70	58	19	60
2.	Andini	95	52	90	72	77
3.	Anisa	89	46	91	22	62
4.	Arkhan	95	49	90	60	73
5.	Balqis	95	48	91	35	67
6.	Chicco	95	47	91	30	66
7.	Clarisa	95	58	91	61	76
8.	Faiz	95	47	89	23	63
9.	Felicia	95	46	86	40	67
10.	Friska	95	47	91	45	69
11.	Jamayca	99	37	83	17	59
12.	Jonea	91	62	90	61	76
13.	Khaliza	95	47	99	33	68
14.	Ledy	95	52	91	56	73
15.	Lirman	86	52	95	14	62
16.	Luthfi	86	56	91	36	67
17.	Natassa	95	42	91	33	65
18.	Orin	91	66	91	57	76
19.	Pratiwi	95	58	90	40	71
20.	Sabar	87	41	76	10	53
21.	Sholayka	95	43	91	56	71
22.	Wildan	88	53	90	11	60
23.	Willy	95	44	91	16	61
24.	Yahsya	95	44	91	29	65

Figure 2. Student Performance in English Language Skills

The score distribution shows significant variation among students across all four skills. The analysis below discusses each skill individually, supported by charts and mean score comparisons.

Listening

Listening comprehension was a relative strength for most students. The average listening score was 93.5, with a highest score of 99 and lowest of 86. Most students performed well on tasks involving direct information and short dialogues. The students were guided with clear audio input and visual support in some tasks.

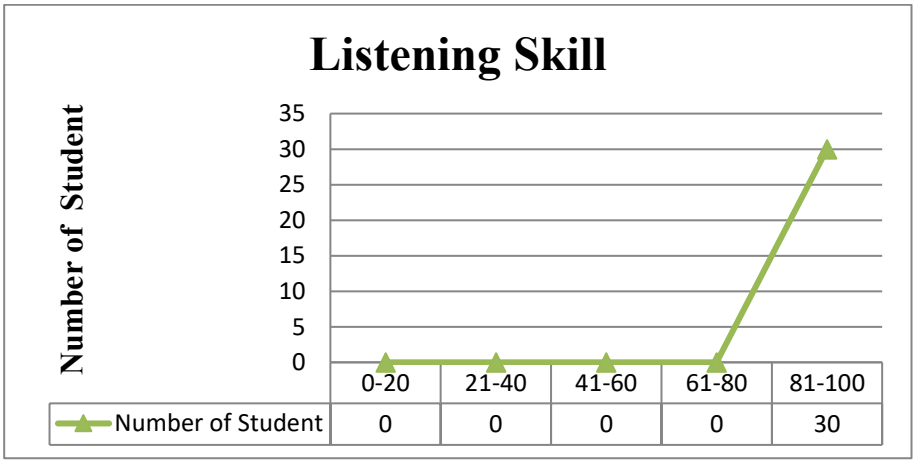

Figure 3. Bar Chart of Listening Scores

Reading

The reading section showed more variability, with scores ranging from 37 to 70, and an average of 49.5. Students had difficulty interpreting main ideas and making inferences. Those with higher reading scores demonstrated stronger vocabulary knowledge and scanning abilities.

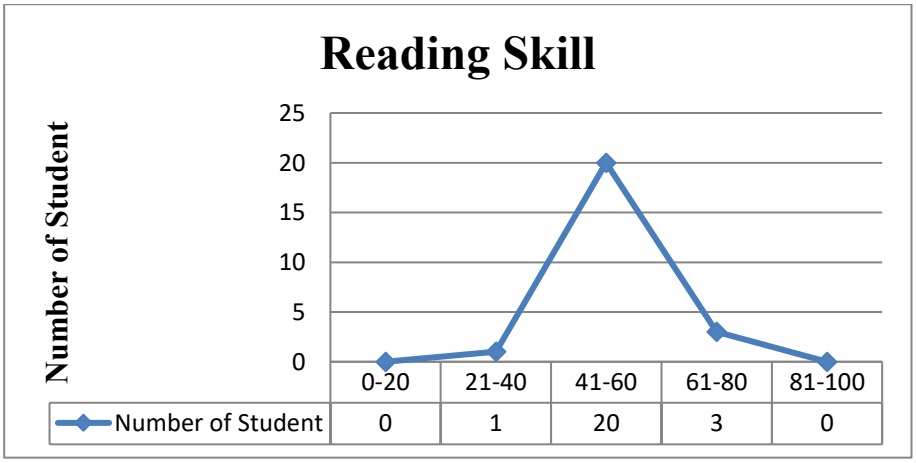

Figure 4. Bar Chart of Listening Scores

Speaking

The speaking component was scored using a structured rubric. Students performed well, with a mean score of 90.25, highest score 99, and lowest score 58. Fluency and pronunciation were generally good, although vocabulary choice and grammar accuracy varied.

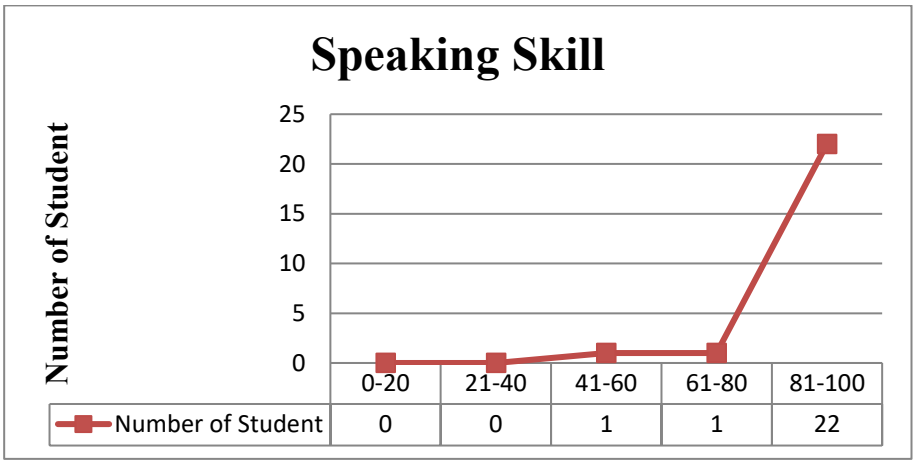

Figure 5. Bar Chart of Speaking Scores

Writing

Writing was the weakest skill. The average score was 41.08, with a highest score of 61 and lowest score of 10. The most common problems included lack of coherence, poor sentence structure, and limited vocabulary. However, some students showed good organization and relevance in their ideas.

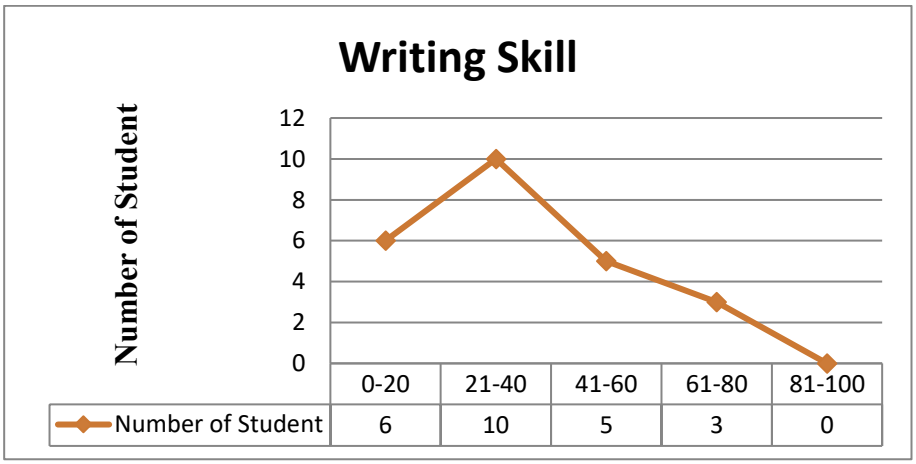

Figure 6. Bar Chart of Writing Scores

Overall Performance

Students showed strong performance in listening and speaking, moderate results in

reading, and relatively weak writing skills. The performance gap suggests a need for enhanced instructional focus on reading comprehension and writing composition.

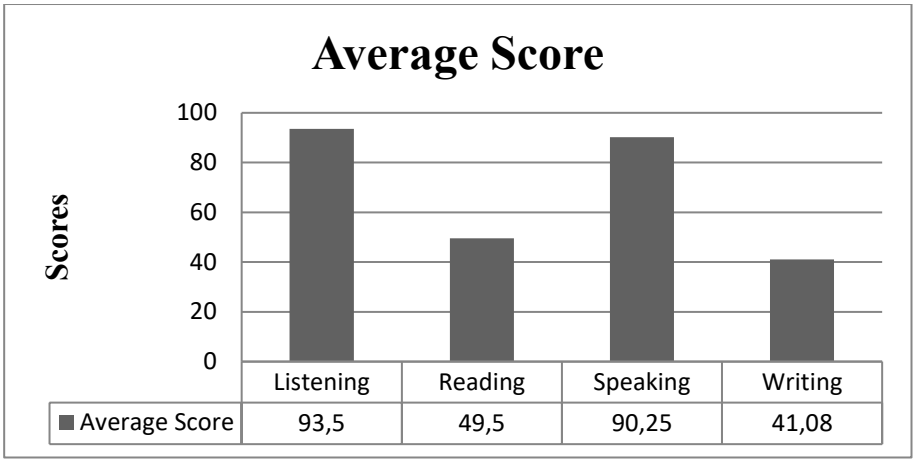

Figure 7. Overall Student Performance

Assessment Evaluation Based on Five Principles

The test was designed and evaluated using the five principles of language assessment: validity, reliability, practicality, authenticity, and washback (Brown & Abeywickrama, 2019).

Validity

The test showed high content validity because it was developed based on indicators from Kurikulum Merdeka. Each skill was assessed using relevant tasks: for example, listening involved short conversations, writing involved descriptive paragraphs, and speaking required oral responses to prompts.

Reliability

Reliability was ensured through a double-rater scoring system. Two assessors evaluated each student's speaking and writing performance using standardized rubrics. The following table presents a comparison of scores from both raters.

No	Name S/V	Listening S/V	Reading S/V	Speaking S/V	Writing S/V	Total S/V
1	Alfahri	92/92	70/71	58/57	19/20	60/60
2.	Andini	95/96	52/52	90/90	72/72	77/77
3.	Anisa	89/88	46/46	91/92	22/22	62/62
4.	Arkhan	95/95	49/48	90/91	60/60	73/73
5.	Balqis	95/95	48/48	91/91	35/34	67/67

6.	Chicco	95/95	47/47	91/91	30/30	66/66
7.	Clarisa	95/94	58/59	91/90	61/61	76/76
8.	Faiz	95/94	47/46	89/88	23/23	63/63
9.	Felicia	95/96	46/47	86/87	40/41	67/68
10.	Friska	95/95	47/47	91/91	45/45	69/70
11.	Jamayca	99/98	37/38	83/83	17/18	59/59
12.	Jonea	91/92	62/61	90/90	61/60	76/76
13.	Khaliza	95/96	47/48	99/99	33/34	68/69
14.	Ledy	95/94	52/51	91/90	56/56	73/73
15.	Lirman	86/86	52/53	95/95	14/14	62/62
16.	Luthfi	86/85	56/55	91/90	36/36	67/66
17.	Natassa	95/95	42/43	91/92	33/34	65/66
18.	Orin	91/91	66/66	91/90	57/57	76/76
19.	Pratiwi	95/95	58/57	90/90	40/41	71/71
20.	Sabar	87/86	41/41	76/75	10/11	53/53
21.	Sholayka	95/96	43/44	91/91	56/57	71/72
22.	Wildan	88/88	53/52	90/91	11/12	60/61
23.	Willy	95/95	44/44	91/90	16/16	61/61
24.	Yahsya	95/94	44/43	91/91	29/30	65/64

Figure 8. Table comparing two raters' scores

Practicality

The test was administered during regular class hours using available resources: speakers for listening, worksheets for reading and writing, and live oral interviews for speaking. It did not require excessive time or budget, confirming high practicality.

Authenticity

Task authenticity was achieved by incorporating real-world and curriculum-based materials. Listening tasks included everyday conversations; writing topics required students to talk about their family; and speaking prompts asked them to describe routines or preferences.

Washback

Positive washback was observed. Students became more aware of their language

ability through detailed feedback. Teachers also used the results to identify areas for further improvement, especially in writing and reading. The test promoted reflection and instructional planning.

CONCLUSION

This study evaluated the English proficiency of Grade VII-2 students at SMP Negeri 2 Siantar and the quality of the assessment instrument based on five core language assessment principles. Results showed varied performance across skills, highlighting strengths in speaking and listening, and weaknesses in reading and writing. The validated and reliable test proved practical, authentic, and capable of generating positive washback. These findings advance classroom-based assessment practices by offering a model that aligns with curriculum standards while supporting instructional improvement. Future research may extend this approach to larger samples and other educational contexts.

REFERENCES

Bachman, L. F., & Palmer, A. S. 1996. Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press.

Brown, H. D. 2004. "Language Assessment: Principles and Classroom Practices. Longman.

Brown, H. D., & Abeywickrama, P. 2010. Language Assessment: Principles and Classroom Practices. Pearson Education.

Douglas, D. 2010. Understanding Language Testing. New York: Routledge.

Gipps, C. 2011. Beyond Testing (Classic Edition): Towards a Theory of Educational Assessment. Routledge.

Jack C. Richards, R. W. 2014. Longman Dictionary of Language Teaching and Applied Linguistics. Routledge.

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia. 2022. Longman Dictionary of Language Teaching and Applied Linguistics. Routledge.

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia. 2022. Capaian pembelajaran jenjang SMP dalam Kurikulum Merdeka. Retrieved Retrieved from https://kurikulum.kemdikbud.go.id.