MIKAILALSYS
Journal of Multidisciplinary Sciences

e-ISSN : 2987-2286
p-ISSN : 2987-3924

Index: Harvard, Boston, Sydney, Copernicus, Sinta 4, Scilit, Semantic, Google, etc

https://doi.org/10.58578/mikailalsys.v4i2.11264

The Quality of English Test Items for Senior High School Students: Validity, Cognitive Level, and Language Skills Analysis

Hevriani Sevrika[1], Yenni Rozimela[2], Syahrul R[3]
[1]Universitas Nahdlatul Ulama Sumatera Barat, Indonesia
[2,3]Universitas Negeri Padang, Indonesia
sevrikariecha@gmail.com; yennirozimela@fbs.unp.ac.id

Article Info:

Submitted:	Revised:	Accepted:	Published:
May 31, 2026	Jun 28, 2026	Jul 10, 2026	Jul 15, 2026

Abstract

Although educational assessment has received considerable scholarly attention, comprehensive evaluations of teacher-made English examinations integrating validity, cognitive demand, and language-skill representation remain limited. This study evaluated the quality of a senior high school English summative examination by examining these three dimensions simultaneously. A quantitative descriptive design was employed to analyze all 50 multiple-choice and essay items included in the examination through total sampling. Data were obtained through documentation. Content validity was assessed based on curriculum alignment, empirical validity was examined using Pearson product–moment correlation, cognitive levels were classified according to Bloom's Revised Taxonomy (C1–C6), and language-skill representation was analyzed using percentage-based descriptive statistics. The findings showed that 38 items (76%) were valid, whereas 12 items (24%) were invalid because of ambiguous wording, poorly constructed distractors, and insufficient alignment with instructional objectives. The cognitive-level distribution was dominated by lower-order

Mikailalsys is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

thinking skills, particularly understanding (C2; 32%) and remembering (C1; 28%). Higher-order thinking skills were substantially underrepresented, comprising analyzing (C4; 12%), evaluating (C5; 6%), and creating (C6; 2%). The examination also emphasized receptive skills and structural knowledge, with reading comprehension accounting for 48% of the items, followed by grammar (20%) and vocabulary (14%). Productive and communicative skills received considerably less coverage, including writing (10%), listening (6%), and speaking (2%). These findings demonstrate that acceptable item validity alone does not ensure a cognitively balanced and communicatively authentic language assessment. This study contributes to the English language assessment literature by providing a multidimensional evaluation of teacher-made examinations in secondary education. It also highlights the need for teachers, schools, and policymakers to conduct continuous item validation, strengthen assessment literacy, increase higher-order cognitive demands, and achieve a more balanced representation of receptive, productive, and communicative language skills.

Keywords: Bloom's Revised Taxonomy; Empirical Validity; English Language Assessment; Language Skill Representation; Teacher-Made Tests

INTRODUCTION

Assessment plays an important role in English language teaching because it provides evidence of students' language competence, learning achievement, and communicative performance. In senior high schools, English examinations are commonly used to evaluate students' mastery of reading, writing, listening, speaking, grammar, and vocabulary. The quality of assessment instruments strongly influences the accuracy of learning evaluation because poorly constructed test items may fail to measure students' actual abilities and produce misleading interpretations of achievement. In educational assessment, validity and reliability are considered fundamental principles because inaccurate instruments weaken the credibility of measurement results. Sürücü and Maslakc (2020), Fitriani. N et al, (2025), and Cook & Hatala (2016) emphasized that high-quality assessment instruments are essential to ensure accurate educational decision-making, while Winna.,and Sabarun (2023), and Gebremariam (2023) argued that it enables teachers and students to receive insightful feedback while also tracking the student's development over time.

The quality of English test items has become a major concern in language assessment research, particularly in secondary education. Recent studies revealed that many teacher-made English examinations still contain weak items that do not meet acceptable

psychometric standards. Dewi et al (2023), Syukur (2025), and Lester & Lee (2012) found that several English reading comprehension items failed to satisfy Rasch model requirements because of inappropriate item difficulty and weak item fit. Similarly, Kusumaningrum (2025) reported that many English test items developed by teachers demonstrated low discrimination power and inconsistent validity indices. Sakinah, (2021) also found that some English achievement test items required substantial revision because they did not adequately represent instructional objectives and curriculum standards. These findings indicate that test quality remains a significant issue in English language assessment and requires continuous evaluation to ensure that examinations accurately reflect students' competencies.

One of the most important indicators of test quality is validity. Validity refers to the extent to which a test measures what it is intended to measure. Tavakol & Wetzel (2020) explained that construct validity determines whether an instrument accurately represents the theoretical concept underlying the assessment. In language testing, validity includes content validity, construct validity, and criterion-related validity. Modern language assessment increasingly emphasizes evidence-based validation procedures to ensure fairness and meaningful interpretation of test scores. Im et al., (2019) argued that validity in language assessment should integrate theoretical frameworks, contextual testing conditions, and score interpretation. Likewise, Kim (2020) highlighted that validity in English proficiency assessment is closely related to educational consequences, curriculum implementation, and policy decisions. Therefore, evaluating validity is essential to determine whether English test items appropriately measure communicative language competence and learning objectives.

Besides validity, cognitive level is another important aspect in evaluating English test quality. Bloom's Revised Taxonomy classifies cognitive processes into remembering, understanding, applying, analyzing, evaluating, and creating (Saputra et al., 2021). Educational assessments are expected to measure not only lower-order thinking skills (LOTS) but also higher-order thinking skills (HOTS) because modern education emphasizes critical thinking, problem-solving, and analytical abilities. However, several studies showed that English examinations are still dominated by lower cognitive levels. Mutiah (2022) found that English summative tests in Indonesian schools primarily focused on remembering and understanding skills. Similarly, Astrid, (2022)) reported that HOTS-oriented English questions remained limited despite curriculum demands emphasizing critical and analytical thinking. This condition may reduce students' opportunities to develop evaluative reasoning and communicative problem-solving skills through language learning activities.

The dominance of lower cognitive levels in English assessment also affects students' intellectual engagement during the learning process. Brookhart (2010) explained that higher-order thinking assessment encourages students to analyze information critically, synthesize ideas, and apply knowledge in unfamiliar situations. In addition, Fitriani et al., (2023) found that many English teachers experienced difficulties in constructing HOTS-based questions because of limited assessment training and insufficient understanding of cognitive frameworks. Boral et al., (2025) similarly revealed that English examinations in secondary education frequently emphasized factual recall rather than analytical language use. Recent studies additionally showed that cognitive complexity significantly influences students' response processes and test performance. Jaya et al., (2026) found that cognitively complex English reading items required deeper comprehension and critical reasoning compared to literal comprehension questions. These findings suggest that cognitive level analysis is necessary to determine whether English test items sufficiently support higher-order thinking development.

Another important component in evaluating English test quality is language skill representation. English learning at the senior high school level aims to develop communicative competence through integrated language skills, including listening, speaking, reading, and writing. Nevertheless, many English examinations focus heavily on reading comprehension and grammar while neglecting productive and communicative skills. Delfi & Purwanti (2022) noted that English assessment practices in Indonesia often fail to proportionally represent communicative language skills. In line with this issue, Hartanti (2022) found that school English examinations were still dominated by reading-based questions with limited assessment of listening and speaking abilities. This imbalance potentially weakens the accuracy of proficiency measurement because communicative competence cannot be fully represented through grammar and reading tasks alone.

The importance of balanced language skill assessment is also supported by communicative language assessment theories. Bachman and Palmer (2010) in Rashed et al., (2018), and Congmin (2023) argued that language assessment should integrate communicative language ability, strategic competence, and contextual language use. Furthermore, recent studies emphasized that communicative assessment contributes positively to students' interactional competence and language confidence. Nurhayati et al. (2023) in Yusuf et al., (2025), and Fitrawati & Safitri (2021) reported that communicative English assessment improved students' speaking confidence and classroom interaction, while

Adem & Berkessa (2022), found that HOTS-based English assessment positively influenced analytical reading skills and classroom engagement. These findings indicate that high-quality English assessment should evaluate both communicative competence and cognitive processes in a balanced manner.

Technological development has also transformed English language assessment practices in recent years. Digital assessment systems, computer-based testing, and artificial intelligence are increasingly used in educational evaluation. (Emmiyati et al., 2024) explained that internet-based English proficiency tests require strong validity, reliability, and practicality to maintain assessment quality and fairness. Similarly, Säuberli & Clematide (2024) found that artificial intelligence could generate English reading comprehension questions with relatively acceptable quality, although issues related to construct representation and linguistic appropriateness remained challenging. Li et al. (2026) in Marie et al., (2025) also reported that AI-generated English reading tests demonstrated promising psychometric characteristics but still required human evaluation to ensure validity and fairness. These developments demonstrate that English language assessment is becoming increasingly complex and requires more comprehensive evaluation frameworks.

Although previous studies have discussed validity, cognitive level, psychometric quality, communicative assessment, and language skill representation, most studies analyzed these aspects separately. Many studies focused only on item validity, HOTS implementation, or psychometric characteristics without comprehensively examining whether English test items simultaneously represent appropriate validity, balanced cognitive levels, and proportional language skill coverage within the same assessment instrument. As a result, a comprehensive understanding of how English test items reflect communicative competence and cognitive complexity in senior high school contexts remains limited, particularly in Indonesian secondary education.

Based on these issues, this study focuses on analyzing the quality of English test items for senior high school students in terms of validity, cognitive level, and language skill representation. The study aims to examine whether English examination items adequately measure intended learning objectives, represent balanced cognitive processes, and proportionally assess communicative language skills. The findings of this study are expected to contribute theoretically to language assessment research and practically to support English

teachers in developing more valid, balanced, and cognitively appropriate examinations that promote communicative competence and higher-order thinking skills.

METHODS

Research Design

This study employed a descriptive quantitative approach. Descriptive quantitative research is appropriate because the study aimed to describe and evaluate the quality of English test items without manipulating variables or testing causal relationships. Specifically, the study focused on examining three dimensions of test quality: item validity, cognitive level distribution, and language skill representation in senior high school English

The study adopted a document analysis design within a descriptive quantitative framework. This design enabled the researchers to systematically evaluate existing English examination items based on predetermined analytical criteria. The analysis was conducted in three stages. First, the validity of the test items was examined to determine whether the items appropriately measured the intended learning objectives. Second, the cognitive levels of the items were identified using Bloom's Revised Taxonomy. Third, the distribution of language skills represented in the examination was analyzed to determine whether the assessment proportionally reflected communicative language competence. This design ensured alignment between the research objectives, the documentary data source, and the analytical procedures employed.

Participants

The participants of this study were not human respondents but English summative test items administered to senior high school students. The sample consisted of all items contained in the selected English summative examinations, including both multiple-choice and essay questions. The test items were selected using a purposive sampling technique because they met specific criteria relevant to the objectives of the study. The criteria included: (1) the examinations were officially administered at the senior high school level, (2) the tests were designed based on the applicable English curriculum, and (3) the complete testing documents, including test papers, answer keys, and learning indicators, were available for analysis. Purposive sampling was considered appropriate because it was allowed the

researchers to obtain information-rich documents capable of providing comprehensive evidence regarding the quality of English assessment practices.

Research Instrument

The primary instrument used in this study was an item analysis checklist developed based on educational assessment theories, Bloom's Revised Taxonomy, and communicative language assessment principles. The checklist consisted of three analytical components. First, validity indicators were used to examine the correspondence between test items, curriculum objectives, learning indicators, and the constructs intended to be measured. Second, cognitive levels were categorized according to Bloom's Revised Taxonomy proposed by Anderson and Krathwohl (2001), namely remembering (C1), understanding (C2), applying (C3), analyzing (C4), evaluating (C5), and creating (C6). Third, language skill representation was classified into reading, writing, listening, speaking, grammar, and vocabulary categories.

Data Collection Procedure

Data were collected using a documentation technique. The researchers obtained English summative examination papers, answer keys, curriculum documents, and relevant instructional indicators from the selected schools. Subsequently, each test item was coded, classified, and entered into the analysis checklist according to the predetermined categories. To enhance the trustworthiness of the classification process, the analytical framework was developed based on established theories and curriculum references.

Data Analysis

The collected data were analyzed using descriptive quantitative techniques. Content validity was examined by analyzing the alignment between the test items, curriculum objectives, and learning indicators. Empirical validity of the test items was calculated using the Pearson Product-Moment correlation coefficient as follows:

$$r_{xy} = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 (\sum X)^2][N \sum Y^2 (\sum Y)^2]}}$$

where (r_{xy}) represents the item validity coefficient, (N) is the number of respondents, (X) is the item score, and (Y) is the total score.

The distribution of cognitive levels and language skills was analyzed by calculating frequencies and percentages using the following formula:

$$P = \frac{F}{N} \times 100\%$$

where (P) represents the percentage, (F) denotes the frequency of occurrence in each category, and (N) refers to the total number of test items analyzed.

The results of the validity analysis, cognitive level classification, and language skill distribution were interpreted descriptively to determine the overall quality of the English examination items and to identify whether the tests reflected balanced cognitive processes and proportional representation of communicative language skills.

RESULTS

This section presents the findings of the study related to the quality of English test items for senior high school students. The analysis focused on three major aspects: validity, cognitive level, and language skill representation. The analyzed test consisted of 50 English test items, including multiple-choice and essay questions. The findings are presented quantitatively and interpreted descriptively to explain the strengths and weaknesses of the examination.

Validity of English Test Items

Validity analysis was conducted using the Pearson Product-Moment correlation formula to determine whether each test item accurately measured students' English competence. The obtained correlation coefficient rcount of each item was compared with the critical value rtable Items with rcount > r_{table} were categorized as valid, while items with rcount < r_{table} were categorized as invalid.

Based on the analysis, 38 items were classified as valid and 12 items were categorized as invalid.

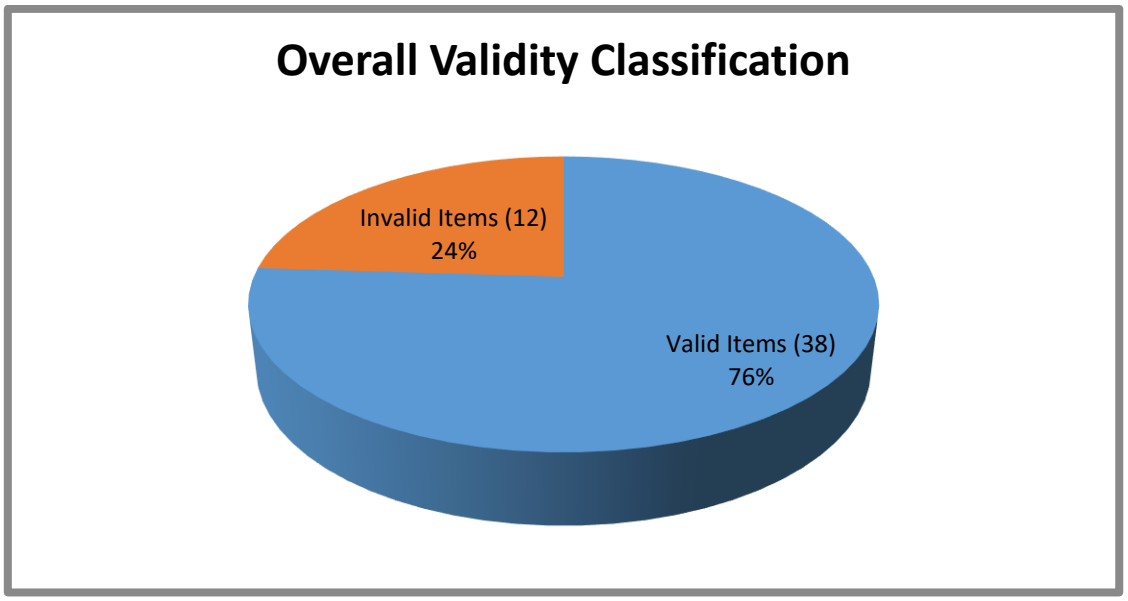


Figure 1. Validity of English Test Items

The results indicate that most test items fulfilled acceptable validity standards. The high percentage of valid items suggests that the examination generally measured the intended English learning objectives. However, the existence of invalid items indicates that several questions still contained weaknesses in construction, language clarity, or content alignment.

Table 1. Valid Test Items

Item Number	r-count	r-table	Category
1	0.52	0.30	Valid
2	0.47	0.30	Valid
3	0.63	0.30	Valid
4	0.58	0.30	Valid
5	0.44	0.30	Valid
6	0.67	0.30	Valid
7	0.61	0.30	Valid
8	0.49	0.30	Valid
10	0.55	0.30	Valid
11	0.43	0.30	Valid
12	0.70	0.30	Valid
13	0.51	0.30	Valid
14	0.39	0.30	Valid
15	0.46	0.30	Valid
17	0.59	0.30	Valid
18	0.65	0.30	Valid

Item Number	r-count	r-table	Category
19	0.42	0.30	Valid
20	0.48	0.30	Valid
21	0.56	0.30	Valid
22	0.41	0.30	Valid
24	0.60	0.30	Valid
25	0.45	0.30	Valid
26	0.53	0.30	Valid
27	0.50	0.30	Valid
28	0.62	0.30	Valid
29	0.40	0.30	Valid
31	0.57	0.30	Valid
32	0.66	0.30	Valid
33	0.54	0.30	Valid
35	0.49	0.30	Valid
36	0.71	0.30	Valid
38	0.58	0.30	Valid
39	0.47	0.30	Valid
41	0.61	0.30	Valid
43	0.64	0.30	Valid
45	0.55	0.30	Valid
47	0.46	0.30	Valid
49	0.59	0.30	Valid

The valid items generally showed good alignment with curriculum objectives and learning indicators. These items effectively measured students' reading comprehension, grammar understanding, vocabulary mastery, and writing ability. Most valid questions used clear instructions, understandable sentence structures, and appropriate distractors.

For example, several valid reading comprehension items required students to identify the main idea, infer meaning from context, and determine specific information from texts. Grammar items that were categorized as valid successfully measured students' understanding of tense usage, sentence structure, and contextual grammar application. Additionally, the valid items demonstrated appropriate difficulty levels. Students with high English proficiency tended to answer these questions correctly, while lower-performing students showed

different response patterns. This indicates that the valid items possessed acceptable discrimination power.

The findings support Brown and Abeywickrama (2019), who explained that valid language test items should measure the intended language constructs clearly and consistently. The findings also align with Dewi et al. (2023), who emphasized that valid English test items contribute significantly to accurate measurement of students' language proficiency.

Table 2. Invalid Test Items

Item Number	r-count	r-table	Category
9	0.21	0.30	Invalid
16	0.18	0.30	Invalid
23	0.24	0.30	Invalid
30	0.19	0.30	Invalid
34	0.15	0.30	Invalid
37	0.27	0.30	Invalid
40	0.20	0.30	Invalid
42	0.16	0.30	Invalid
44	0.25	0.30	Invalid
46	0.11	0.30	Invalid
48	0.23	0.30	Invalid
50	0.17	0.30	Invalid

The invalid items indicated several weaknesses in item construction and assessment quality. Some items contained ambiguous wording that confused students and caused inconsistent responses. Other items lacked alignment with learning indicators and measured content that was not sufficiently emphasized during classroom instruction.

Several invalid grammar items also contained distractors that were too easy or obviously incorrect. As a result, the questions failed to discriminate effectively between high-achieving and low-achieving students. Some reading comprehension items provided unclear passages or questions with multiple possible interpretations, reducing the validity of students' responses. For example, Item 34 showed the lowest validity coefficient $r = 0.15$. This item required students to identify implicit meaning from a short text, but the question stem was unclear and did not provide sufficient contextual clues. Similarly, Item 46 demonstrated weak correlation because the distractors were poorly constructed and did not function effectively.

The existence of invalid items suggests that the English test still requires revision and systematic item analysis before implementation. According to Hamzah et al. (2020), invalid items frequently occur because teachers often develop examinations without psychometric evaluation procedures. Rahmawati and Ertin (2019) also argued that teacher-made English tests commonly face problems related to ambiguous language use and weak distractor quality.

The findings imply that English teachers should conduct item validation processes regularly to improve assessment quality. Invalid items should either be revised or removed from future examinations to ensure that test results accurately represent students' English competence.

Cognitive Level Analysis

The cognitive levels of the English test items were analyzed using Bloom's Revised Taxonomy. The findings showed that the examination was dominated by lower-order thinking skills (LOTS), particularly understanding (C2) and remembering (C1). Meanwhile, higher-order thinking skills (HOTS), such as analyzing (C4), evaluating (C5), and creating (C6), appeared in smaller proportions. This indicates that most questions focused on memorization and comprehension rather than critical and creative thinking.

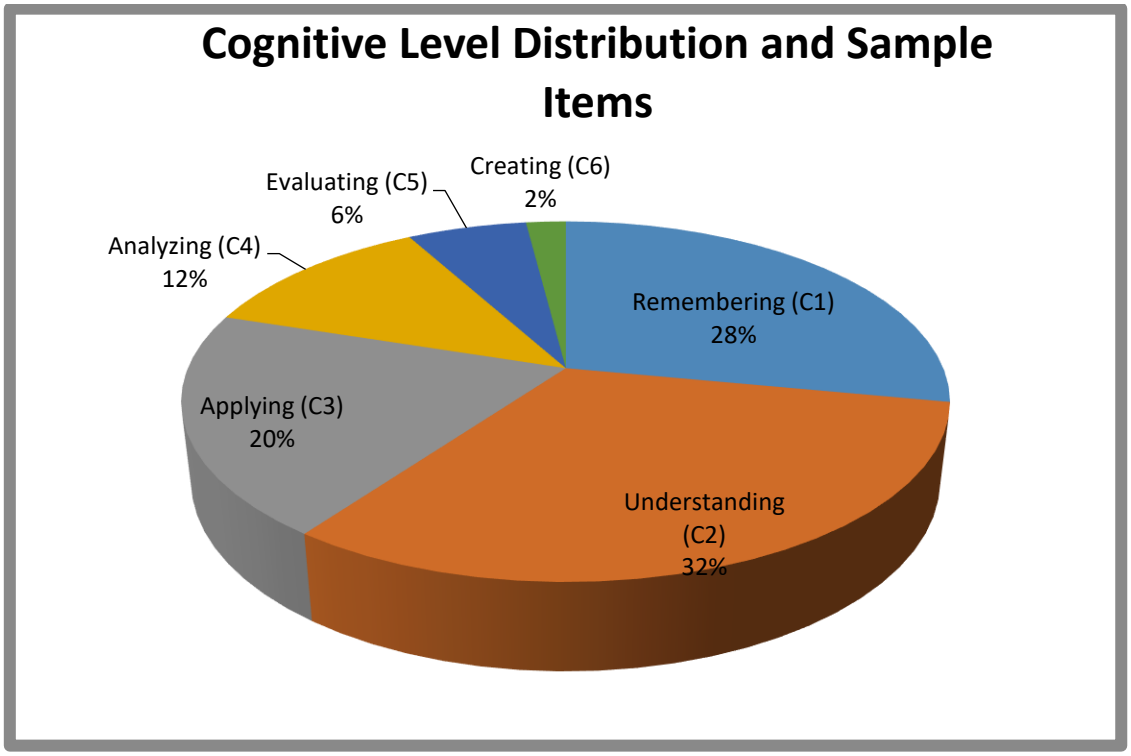


Figure 2. Cognitive Level Distribution and Sample Items

The findings indicate that lower-order thinking skills (LOTS) dominated the examination. The understanding category (C2) became the highest cognitive level with 32%, followed by remembering (C1) with 28%. Most items required students to identify explicit information, recall vocabulary meanings, and understand grammar usage. Meanwhile, higher-order thinking skills (HOTS), including analyzing (C4), evaluating (C5), and creating (C6), only represented 20% of the total items.

The analysis also revealed that C1 items mainly focused on vocabulary memorization and grammar recall. C2 questions emphasized reading comprehension and interpretation of explicit information from texts. Meanwhile, C3 items required students to apply grammar rules and functional expressions in contextual situations. In contrast, HOTS-oriented items appeared in limited numbers. C4 items required students to analyze relationships between ideas and identify text organization. C5 questions focused on evaluating arguments and expressing judgments logically. Only one item reached the C6 level because students were asked to produce original written language through paragraph writing activities.

These findings suggest that the English examination still emphasized memorization and comprehension rather than critical thinking and communicative production. This result supports Zulaiha and Mulyono (2022), who found that English assessments in Indonesian schools are still dominated by LOTS-oriented questions. Therefore, English teachers should increase the number of HOTS-based questions to promote analytical, evaluative, and creative language use among students.

Language Skill Representation

The analysis of language skill representation was conducted to identify how the English examination distributed assessment across language skills and language components. The findings revealed that the examination was dominated by reading and grammar questions, while productive skills such as speaking and writing received minimal attention.

Table 4. Distribution of Language Skills in English Test Items

Language Skill	Frequency	Percentage	Item Numbers	Sample Item
Reading	24	48%	1–24	"What is the main idea of the passage?"
Grammar	10	20%	25–34	"If I ___ enough money, I will buy a new laptop."
Vocabulary	7	14%	35–41	"What is the synonym of 'beautiful'?"

Language Skill	Frequency	Percentage	Item Numbers	Sample Item
Writing	5	10%	42–46	"Arrange the sentences into a good paragraph."
Listening	3	6%	47–49	"What information did the speaker mention?"
Speaking	1	2%	50	"Create a short dialogue about asking and giving directions."
Total	50	100%		

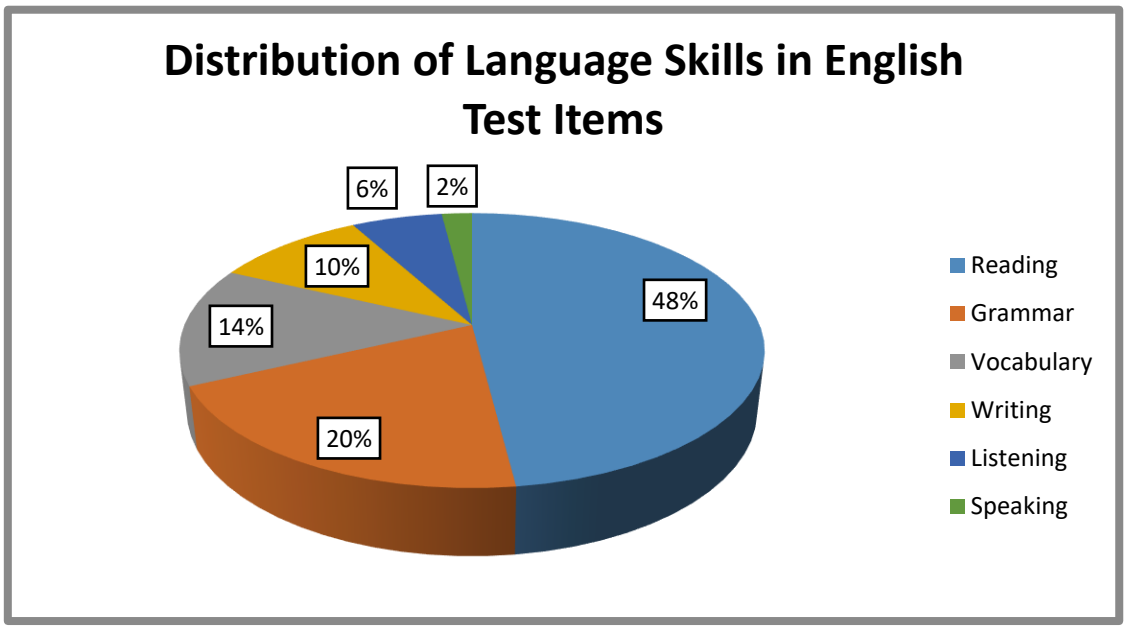


Figure 3. Distribution of Language Skills in English Test Items

The findings demonstrate that reading skill became the dominant component in the examination, representing 48% of the total items. Most reading questions focused on identifying main ideas, finding detailed information, interpreting vocabulary in context, and understanding text structure. This dominance indicates that the examination strongly emphasized receptive language skills, particularly text comprehension.

The large proportion of reading items reflects the common practice in English assessment at the secondary school level, where reading comprehension is often prioritized because it is easier to administer and score objectively. However, excessive focus on reading may reduce opportunities for students to demonstrate productive communicative competence.

DISCUSSION

The findings of this study indicate that the quality of senior high school English test items was moderate. Although most items met acceptable validity criteria, important weaknesses remained in the distribution of cognitive levels and the representation of language skills. These findings address the objectives of the study, which sought to examine whether English examinations adequately measure intended learning outcomes, reflect balanced cognitive processes, and proportionally assess communicative language competence.

The validity analysis revealed that 76% of the test items were categorized as valid, whereas 24% were classified as invalid. This finding suggests that the majority of the examination items were able to measure the intended constructs and learning objectives appropriately. Valid items generally demonstrated alignment with curriculum indicators, clear wording, and appropriate grammatical structures. However, the existence of invalid items indicates that some questions failed to function effectively as measurement tools. Ambiguous wording, ineffective distractors, and unclear contexts may have interfered with students' interpretation of the items, thereby reducing the accuracy of score interpretation.

The findings further demonstrated that the examination was dominated by lower-order thinking skills (LOTS), particularly remembering (C1) and understanding (C2), which together accounted for 60% of the total items. This pattern indicates that the examination primarily assessed students' ability to recall information and comprehend explicit meanings rather than engage in analytical, evaluative, or creative thinking. Consequently, students had limited opportunities to demonstrate critical reasoning, problem-solving abilities, and deeper language processing. The finding is consistent with the objective of the study to determine whether English assessments support the development of higher-order thinking skills.

Another important finding concerns the imbalance in language skill representation. Reading comprehension constituted the largest proportion of the examination items, followed by grammar and vocabulary questions, whereas listening and speaking were minimally represented. This distribution suggests that the examination emphasized receptive language skills while providing insufficient opportunities to assess productive and interactive communication. As a result, the tests only partially reflected students' communicative competence, despite communicative ability being one of the principal goals of English language learning.

The study also identified a close relationship between cognitive level and language skill representation. Most reading comprehension questions were classified at the remembering and understanding levels, indicating that even reading activities largely focused on surface-level comprehension. Students were primarily required to identify explicit information and determine vocabulary meanings rather than interpret implied meanings, evaluate arguments, or synthesize ideas across texts. These findings suggest that the assessment had not fully supported the development of critical literacy skills required in twenty-first-century education.

The finding that most test items demonstrated acceptable validity is consistent with the view of Tarigan et al., (2025), who argued that effective language assessments should accurately represent instructional objectives and provide dependable evidence of learners' abilities. Nevertheless, the presence of invalid items supports the findings of Effendi & Mayuni (2022), who reported that teacher-made English tests frequently lacked systematic validation procedures and often contained ambiguous language structures. Similarly, Suilistio (2023) argued that limited assessment literacy among teachers commonly contributes to poorly functioning items and inadequate discrimination power. Therefore, the current findings reinforce concerns regarding teachers' preparedness in developing high-quality assessment instruments.

The dominance of LOTS-oriented questions is also consistent with previous studies conducted in Indonesian educational contexts. Mutiah (2022) found that English summative assessments predominantly measured remembering and understanding skills, while Astrid (2022) reported that HOTS-oriented questions remained limited despite curriculum expectations promoting critical and analytical thinking. Brookhart (2010) further emphasized that assessments focused mainly on factual recall are unlikely to foster meaningful learning and higher-order reasoning. The present study extends these findings by demonstrating that the limited implementation of HOTS is not only reflected in cognitive distribution but is also closely associated with the types of language skills being assessed.

The imbalance in language skill representation aligns with the findings of Khoirunnisa & Widodo (2019), who noted that English assessment practices in Indonesia continue to be influenced by traditional grammar-translation and reading-oriented approaches. Similarly, Putriani et al (2025) found that productive skills such as speaking and writing are often neglected because they require more preparation time, subjective scoring

procedures, and additional teacher expertise. The findings also support by Kaçar (2023), who argued that multiple-choice formats, although practical and objective, have limitations in assessing authentic language production and communicative performance.

From the perspective of communicative language assessment, the findings appear inconsistent with the framework proposed by Baharuddin (2024) . According to their model, communicative language ability encompasses grammatical competence, textual competence, pragmatic competence, and strategic competence that should be assessed proportionally. However, the examinations analyzed in this study focused predominantly on grammatical recognition and textual comprehension, providing limited opportunities for students to demonstrate pragmatic and strategic language use. This discrepancy suggests that current assessment practices have not yet fully operationalized the principles of communicative language testing.

The findings contribute theoretically to the growing body of literature on language assessment by demonstrating the importance of evaluating test quality through an integrated framework that simultaneously considers validity, cognitive complexity, and language skill representation. Previous studies have often examined these dimensions separately. The present study highlights how these components interact and collectively influence the extent to which examinations reflect communicative competence.

Practically, the findings provide evidence-based recommendations for English teachers and educational institutions. Teachers should conduct systematic item analyses before administering examinations to identify invalid items, revise ineffective distractors, and improve the clarity of item wording. In addition, professional development programs should strengthen teachers' assessment literacy, particularly in designing HOTS-oriented questions and applying principles of communicative assessment.

Schools and educational policymakers should also encourage balanced assessment practices that proportionally represent both receptive and productive language skills. Integrating speaking and listening tasks into school examinations would provide a more comprehensive evaluation of students' communicative abilities. Furthermore, increasing the proportion of analytical, evaluative, and creative tasks may support the development of critical thinking, problem-solving skills, and deeper language engagement that are increasingly emphasized in twenty-first-century education.

The findings additionally highlight the importance of washback effects in English language teaching. Since teachers frequently align classroom instruction with examination demands, improving examination quality may positively influence instructional practices. Assessments that promote authentic communication and higher-order thinking are more likely to encourage active learning, critical inquiry, and meaningful language use in classroom settings.

Several limitations should be acknowledged when interpreting the findings of this study. First, the study focused exclusively on English summative examination documents from a limited educational context. Consequently, the findings may not be generalized to all senior high schools with different curricular practices and assessment cultures. Second, the study relied primarily on document analysis without incorporating teachers' perspectives, students' experiences, or classroom observations that might provide a deeper understanding of the factors influencing assessment practices. Third, the analysis emphasized validity, cognitive levels, and language skill representation without examining other psychometric characteristics such as reliability, item discrimination, and item difficulty.

Future studies are encouraged to expand the scope of investigation by involving a larger number of schools from diverse geographical contexts to improve the generalizability of findings. Researchers may also adopt mixed-methods approaches that combine quantitative item analysis with qualitative interviews or observations to explore teachers' assessment literacy, institutional constraints, and decision-making processes in test construction. In addition, future research should integrate broader psychometric analyses and investigate the effects of communicative and HOTS-oriented assessments on students' language achievement, critical thinking development, and classroom learning experiences.

Overall, this study demonstrates that although the English examination exhibited acceptable validity in many items, substantial improvements remain necessary in terms of cognitive complexity and communicative skill representation. High-quality English assessments should move beyond measuring students' ability to recall grammatical knowledge and identify factual information. Instead, they should provide meaningful opportunities for students to analyze texts critically, solve communication problems, express ideas creatively, and use English effectively in authentic contexts.

CONCLUSION

This study examined the quality of senior high school English summative test items by simultaneously analyzing their validity, cognitive level distribution, and language skill representation. The findings indicate that, although the examination demonstrated acceptable validity in most items, the overall assessment quality remained only moderate because it did not adequately reflect balanced cognitive demands and communicative language competence. The presence of invalid items suggests that weaknesses in item construction still exist, potentially reducing the accuracy of measurement and score interpretation. In addition, the predominance of lower-order thinking skills indicates that the examination largely emphasized knowledge recall and comprehension rather than fostering analytical, evaluative, and creative thinking. The disproportionate emphasis on reading, grammar, and vocabulary, alongside the minimal representation of listening and speaking, further suggests that the assessment only partially captured the integrated communicative abilities expected in English language learning.

The findings contribute to the field of language assessment by demonstrating the importance of evaluating English examinations through an integrated perspective that considers validity, cognitive complexity, and language skill representation simultaneously. While previous studies have frequently examined these dimensions independently, this study highlights how their interaction shapes the extent to which assessments align with curricular goals and communicative language teaching principles. Practically, the study provides evidence that strengthening teachers' assessment literacy, conducting systematic item analyses, and incorporating more authentic and higher-order assessment tasks are essential steps toward improving the quality of school-based English examinations.

Several recommendations emerge from the present study. Future research is encouraged to involve a larger number of schools and more diverse educational contexts to enhance the transferability of the findings. Researchers may also employ mixed-methods approaches by combining quantitative item analysis with qualitative investigations of teachers' assessment practices and decision-making processes. Furthermore, future studies should incorporate additional psychometric indicators, such as reliability, item discrimination, and difficulty indices, as well as explore the effects of communicative and HOTS-oriented assessments on students' language achievement, critical thinking development, and classroom learning experiences. Such efforts will contribute to a more

scomprehensive understanding of how assessment practices can better support meaningful English language learning and the development of students' communicative competence.

REFERENCES

Adem, H., & Berkessa, M. (2022). A case study of EFL teachers' practice of teaching speaking skills vis-à-vis the principles of communicative language teaching (CLT). *Cogent Education, 9*(1), 2087458. https://doi.org/10.1080/2331186X.2022.2087458

Al Mamari, M. R. H., Al-Mekhlafi, A. M., & Al-Barwani, T. A. (2018). Is communicative language teaching being tested communicatively? An analysis of English tests in Oman. *English Language Teaching, 11*(11), 83–91. https://doi.org/10.5539/elt.v11n11p83

Astrid, A., Hasanah, A., & Syafryadin. (2022). Integrating higher order thinking skills (HOTS) into English language teaching for elementary school students: Teachers' perspectives and challenges. *3L: The Southeast Asian Journal of English Language Studies, 28*(3), 217–230. https://doi.org/10.17576/3L-2022-2803-14

Baharuddin, L. (2024). Implementasi Metode Pembelajaran Berbasis Proyek dalam Pembelajaran Bahasa Inggris di MAN 1 Ternate. *JUANGA: Jurnal Agama dan Ilmu Pengetahuan, 10*(1), 37–48. https://www.e-jurnal.staibabussalamsula.ac.id/index.php/JUANGA/article/view/140

Boral, K., Mandal, S. K., Kumbhakar, M., & Dutta, S. K. (2025). Analysis of higher order thinking skills of intended mathematics learning outcomes in India. *MIER Journal of Educational Studies Trends and Practices, 15*(1), 146–176. https://doi.org/10.52634/mier/2025/v15/i1/2856

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: A primer for simulation and beyond. *Advances in Simulation, 1*, 31. https://doi.org/10.1186/s41077-016-0033-y

Dewi, H. H., Damio, S. M., & Sukarno. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *REID (Research and Evaluation in Education), 9*(1), 24–36. https://doi.org/10.21831/reid.v9i1.53514

Effendi, T., & Mayuni, I. (2022). Examining teacher-made English test in a language school. *LADU: Journal of Languages and Education, 2*(2), 67–76. https://doi.org/10.56724/ladu.v2i2.109

Emmiyati, N., Asnur, S. M., & Ahmad. (2025). Validity, reliability, and practicality of test items in an internet-based English proficiency test for English laboratory. *IJEE (Indonesian Journal of English Education), 11*(2), 197–206. https://doi.org/10.15408/ijee.v11i2.37276

Firdaus, Yusuf, Y. Q., Samad, I. A., & Delavari, H. (2025). Enhancing speaking skills in higher education through Canvas LMS: A study in Banda Aceh. *English Review: Journal of English Education, 13*(2), 407–420. https://doi.org/10.25134/erjee.v13i2.11561

Fitrawati, & Safitri, D. (2021). Students' grammatical errors in essay writing: A pedagogical grammar reflection. *International Journal of Language Education, 5*(2), 74–88. https://doi.org/10.26858/ijole.v5i2.15241

Fitriani, N., Rahmadewi, S., Nurhadi, & Shuhada. (2025). Development and role of measurement tools in educational evaluation. *JMPI: Jurnal Manajemen, Pendidikan dan Pemikiran Islam, 3*(1), 78–97. https://doi.org/10.71305/jmpi.v3i1.233

Fitriani, R., Iskandar, I., & Suseno, M. (2023). The implementation of higher order thinking skills (HOTS) in English language teaching: A case of Indonesian senior high school EFL teachers. *Didaktika: Jurnal Kependidikan, 12*(4), 739–750. https://doi.org/10.58230/27454312.303

Gebremariam, H. T., & Gedamu, A. D. (2023). Primary school teachers' assessment for learning practice for students' learning improvement. *Frontiers in Education, 8*, 1145195. https://doi.org/10.3389/feduc.2023.1145195

Hartanti, T. P. (2022). *Analysis of students' speaking skill at ninth grade in Junior High School 2 Jambi City* [Undergraduate thesis, Batanghari University]. http://repository.unbari.ac.id/1959/

Im, G.-H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia, 9*, 14. https://doi.org/10.1186/s40468-019-0089-4

Jaya, S., Kusmiarti, R., & Paulina, Y. (2026). Developing critical reading behaviors through guided close reading of research articles. *Indonesian Journal of Educational Development, 6*(4), 1569–1587. https://doi.org/10.59672/ijed.v6i4.5698

Jismulatif, Dahnilsyah, Delfi, S., & Purwanti, I. T. (2022). The representation of foreign and Indonesian cultures in English textbooks for junior high school in Indonesia. *Al-Ishlah: Jurnal Pendidikan, 14*(3), 4217–4230. https://doi.org/10.35445/alishlah.v14i1.2016

Kaçar, K. (2023). Commonly used techniques in testing foreign language skills and communicative language teaching. *Sinerji Uluslararası Alan Eğitimi Araştırmaları Dergisi, 4*(2), 68–85. https://doi.org/10.54971/synergy.1381532

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature and Culture, 6*(2), 256–269. https://doi.org/10.30659/e.6.2.256-269

Khoirunnisa, A., & Widodo, E. (2019). Students' difficulties in comprehending narrative text. *TELL: Teaching of English Language and Literature Journal, 7*(2), 65–74. https://doi.org/10.30651/tell.v7i2.3441

Lee, H. L. J. (2012). SMARTies: Using a board game in the English classroom for edutainment and assessment. *Malaysian Journal of ELT Research, 8*(1), 1–35. https://meltajournals.com/index.php/majer/article/view/616

Mutiah, S. D. (2023). *English teachers' assessment literacy and online formative assessment implementation in teaching writing: A mixed-method study* [Thesis, UIN Syarif Hidayatullah Jakarta]. https://repository.uinjkt.ac.id/dspace/handle/123456789/66568

Putriani, Ilham, Rahmaniah, R., & Humaira. (2025). The effect of peer feedback on English writing skills among EFL students. *ELTIN Journal: Journal of English Language Teaching in Indonesia, 13*(2), 433–442. https://e-journal.stkipsiliwangi.ac.id/index.php/eltin/article/view/5991

Ripoll Y Schmitz, L. M., & Sonnleitner, P. (2025). Evaluating AI-generated vs. human-written reading comprehension passages: An expert SWOT analysis and comparative study for an educational large-scale assessment. *Large-Scale Assessments in Education, 13*, 20. https://doi.org/10.1186/s40536-025-00255-w

Saputra, H. T., & Pujiati, H. (2021). Cognitive domain of revised Bloom taxonomy in English student book. *JELT: Journal of English Language and Literature Teaching, 6*(1), 61–66. http://ejournal.unima.ac.id/index.php/jellt/article/view/2816

Säuberli, A., & Clematide, S. (2024). Automatic generation and evaluation of reading comprehension test items with large language models. *Proceedings of the Third Workshop on Tools and Resources for People with REAding DIfficulties*, 22–37. https://aclanthology.org/2024.readi-1.3

Suharyadi, & Kusumaningrum, S. R. (2025). Linking item analysis to teaching and learning: What English test data tell us about teachers' competencies. *Jurnal Pengabdian Masyarakat dan Riset Pendidikan, 4*(2), 9109–9116. https://doi.org/10.31004/jerkin.v4i2.3116

Sulistio, P. H. (2023). Language assessment literacy: Cases of pre-service English teachers. *IJLECR: International Journal of Language Education and Cultural Review, 9*(1), 1–10. https://doi.org/10.21009/ijlecr.v9i1.33032

Sürücü, L., & Maslakçı, A. (2020). Validity and reliability in quantitative research. *Business & Management Studies: An International Journal, 8*(3), 2694–2726. https://doi.org/10.15295/bmij.v8i3.1540

Syukur, B. A., & Nurlaily, A. F. (2025). Rasch model-based evaluation of TOEFL listening items: Analyzing difficulty, discrimination, and fit. *Jurnal Smart, 11*(2), 176–191. https://doi.org/10.52657/js.v11i2.2931

Tarigan, A. F. A., Simanjuntak, P. H. T. M., Sitopu, C. D., Purba, J. H., & Silalahi, D. E. (2025). The implementation of English language assessment based on language assessment and evaluation principles at senior high school. *Jurnal Pendidikan Bahasa, 15*(3), 447–454. https://doi.org/10.37630/jpb.v15i3.3412

Tavakol, M., & Wetzel, A. (2020). Factor analysis: A means for theory and instrument development in support of construct validity. *International Journal of Medical Education, 11*, 245–247. https://doi.org/10.5116/ijme.5f96.0f4a

Winna, & Sabarun. (2023). The language assessment in teaching-learning English. *DIAJAR: Jurnal Pendidikan dan Pembelajaran, 2*(4), 413–419. https://doi.org/10.54259/diajar.v2i4.1894

Wolf, M. K. (Ed.). (2020). *Assessing English language proficiency in U.S. K–12 schools*. Routledge. https://doi.org/10.4324/9780429491689

Zhao, C. (2023). Theorizing the language ability and language assessment. *US-China Foreign Language, 21*(7), 261–266. https://doi.org/10.17265/1539-8080/2023.07.002