IJOMER
(Indonesian Journal of Multidisciplinary Educational Research)
P-ISSN: 2987-7601 E-ISSN: 2987-0968
https://jurnalfaktarbiyah.iainkediri.ac.id/index.php/ijomer
DOI: https://doi.org/10.30762/ijomer.v4i2.9275

An Analysis of the Quality of Multiple-Choice Test Items in the Final Summative Assessment for Grade X Indonesian Language Learning

***Mila Rahmawati[1]**
Universitas Islam Negeri Madura, Indonesia
* Email: 23381072042@student.uinmadura.ac.id
Ririn Fatmawati[2]
Universitas Islam Negeri Madura, Indonesia
Email: 23381072058@student.uinmadura.ac.id
Liana Rochmatul Wachidah[3]
Universitas Islam Negeri Madura, Indonesia
Email: lianarwachidah@uinmadura.ac.id

(*) Email Correspondent: 23381072042@student.uinmadura.ac.id

Abstract

This study aims to evaluate the quality of multiple-choice items in the Final Summative Assessment covering Material Scope or *Asesmen Sumatif Akhir Lingkup Materi* (ASALM) for Grade X Indonesian language, based on the aspects of content, construction, and language. The researcher used a descriptive qualitative approach, analyzing 40 items through a document study of the test items, the item blueprint, the answer key, and other supporting documents. The data were analyzed using the Miles, Huberman, and Saldaña model, which includes data reduction, data display, and conclusion drawing. According to the findings, 80% of the things matched the learning indicators in terms of content, but the material was not yet uniformly distributed since negotiation texts accounted for 40% of the items. In addition, lower-order thinking skills (LOTS) items (60%) continue to make up the majority of the products, while higher-order thinking skills (HOTS) items only make up 30% of them. While weaknesses were still discovered in the arrangement of the answer options, 85% of the items adhere to the writing rules in terms of construction. In term of language, there were a few writing mistakes, erratic use of language, and sentences that might be interpreted in more than one way. Overall, the quality of the items is rather good, but more improvement is still needed in terms of content balance, language usage, item construction, and the percentage of HOTS items to make them more in line with the *Kurikulum Merdeka* (Independent Curriculum).

Keywords: HOTS; Indonesian language; item quality; *Kurikulum Merdeka;* summative assessment

Received:	25-06-2026	Revised:	22-10-2026	Accepted:	31-10-2026

Copyright: © 2022 Fakultas Tarbiyah, Institut Agama Islam Negeri Kediri. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution - ShareAlike 4.0 International License (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

INTRODUCTION

Evaluation is one important part of the education process. It helps show how far learning goals have been reached and becomes a basis for decisions to improve the quality of learning. The success of an evaluation depends a lot on the quality of the tool used. An effective measurement tool must be able to assess students' ability accurately, fairly, and in line with the learning goals that have been set. Therefore, it is essential to consider various factors when developing an assessment tool to ensure that the results accurately represent students' skills (Sudijono, 2017).

In the implementation of *Kurikulum Merdeka*, assessment is an inseparable part of the learning process. One form of assessment used is the Final Summative Assessment covering Material Scope or *Asesmen Sumatif Akhir Lingkup Materi* (ASALM), which is carried out after students finish a certain scope of learning material. This assessment aims to find out how well students have mastered the competencies they have learned. In practice, multiple-choice items are still widely used because they can cover a broad range of material and make scoring and processing of learning results easier. Even so, attention to the use of multiple-choice items in Indonesian language evaluation remains important, because the skills assessed in this subject involve not only the ability to remember, but also understanding, analysis, evaluation, and the application of language in various contexts. This raises a question about how far multiple-choice items are truly effective in measuring Indonesian language competence as a whole. In addition, there is a possibility that the items written measure lower order thinking skills (LOTS) more than higher order thinking skills (HOTS). The construction of the items also needs attention, because item quality is affected by how well the items follows writing rules, such as the clarity of the item stem, the accuracy of the language used, the similarity (homogeneity) of the answer options, and how well the distractors work (Arikunto, 2018).

The researcher's initial observation also indicates the need for careful attention to the construction of multiple-choice items. Based on the researcher's initial observation of the End-of-Year Assessment documents for Grade X Indonesian language at MA Al Falah Pamekasan, it was found that most items were still dominated by questions that measure lower order thinking skills (LOTS), such as remembering definitions, identifying text structure, and recognizing language features. This can be seen in several items that ask students to choose an answer based on information they had directly studied. Meanwhile,

items that measure higher order thinking skills (HOTS), such as analyzing, evaluating, and giving a critical judgment on a text, were still limited. In addition, the researcher also found some signs of weakness in item construction, such as inconsistent writing of answer options and the use of language that could lead to more than one interpretation. Even so, the quality of these items remains important, because the skills assessed in this subject involve not only recall, but also understanding, analysis, evaluation, and the application of language in various contexts. This raises a question about how far multiple-choice items are truly effective in measuring Indonesian language competence as a whole. Analysis of of the content, construction, and language aspects is important in order to find out the quality of the items used in the assessment. Through this analysis, the weaknesses and strengths of each item can be identified more deeply, so they can become a basis for improving the assessment instrument (Kemendikbud, 2020). In addition, a good test must meet certain requirements so it can work as an accurate measurement tool in the learning evaluation process (Arikunto, 2018).

Several studies on item quality analysis have been carried out with different focuses. Syahriandi (2014) analyzed the quality of Indonesian language items based on validity, reliability, difficulty level, discriminating power, and distractor function, while Syahriandi (2017) focused more on a qualitative analysis of the language aspect. Another study by Hariati et al. (2026) focused on a statistical analysis of item quality using the ANATES software. Globally, investigations into multiple-choice questions have examined aspects such as the effectiveness of these questions in assessing what they are intended to and issues related to their phrasing. A comprehensive review of these studies reveals that the majority assess question quality only to a limited extent, either focusing solely on quantitative data or merely on the wording involved. Furthermore, many of these investigations originate from the field of medical education, rather than from assessments involving the Indonesian language.

An effective multiple-choice question is typically evaluated based on three primary dimensions: content, which pertains to its alignment with educational objectives and the cognitive level it aims to evaluate; construction, which ensures the presence of a precise inquiry, a single correct response, and distractors that appear plausible but are incorrect; and language, which guarantees clarity, readability, and the absence of ambiguity. However, there has been a lack of

studies that comprehensively address all three areas in a single review that adheres to the standards of Grade X tests in the Indonesian language.

This research examines the quality of multiple-choice questions present in the Final Summative Assessment for Grade X Indonesian, concentrating on their alignment with the syllabus, structural design, and linguistic clarity. Since it does not analyze student responses, it does not provide statistical data on items. Its primary contribution lies in offering a consolidated method for assessing the quality of test questions that educators can utilize to make enhancements prior to their future application.

Based on the description above, this study aims to analyze the quality of multiple-choice items in the Final Summative Assessment covering Material Scope for Grade X Indonesian language, based on the aspects of content, construction, and language. The results of the study are expected to give a picture of the quality of the items used, and to become a consideration in preparing better assessment instruments in the future.

METHOD

Research Design

Because the goal of the study was to thoroughly characterize and analyze the quality of multiple-choice items according to content, construction, and language, using item-writing guidelines without student response data or empirical item statistics, the researcher selected a descriptive qualitative approach. Descriptive research serves to give an organized picture of the quality of the items used in the Final Summative Assessment covering Material Scope or ASALM for Grade X Indonesian language.

Subject of the Research

The subject of this study was 40 multiple-choice items in the Final Summative Assessment covering Material Scope or ASALM for Grade X Indonesian language. Included in the data sources were various supporting documents concerning the creation of the assessment instrument, the item blueprint, the answer key, and 40 multiple-choice items. The materials were administered to 50 Grade X students at MA Al-Falah Pamekasan, during the 2025/2026 academic year's even semester. The Indonesian language teacher prepared the exam questions, which included sections on poetry, effective sentences and spelling, observation report texts, biographies, *hikayat*, and the

Material Scope of negotiation texts. The item indicators, cognitive levels, and learning goals (*Tujuan Pembelajaran/TP*) were all included in the item blueprint.

Instruments

The researcher collected the data through document study, by analyzing all the files that were the focus of this study. The research instrument was an item analysis sheet prepared based on the principles of writing multiple-choice items. This analysis focused on three main components: the content component, the structure component, and the language component. The content component covers how well the question matches the indicator and the learning goal, as well as the accuracy of the information being assessed. The structure component covers the clarity of the item stem, the consistency of the answer options, and the accuracy in arranging the alternative answers. Meanwhile, the language component covers the correct use of Indonesian, the clarity of the sentences, and how relevant they are to students' stage of development.

Data Analysis

The researcher analyzed the data using the interactive analysis model proposed by (Miles, 2019) which includes the steps of data reduction, data display, and conclusion drawing. Data obtained from the document analysis were grouped according to the aspects of content, structure, and language, then interpreted to assess the quality of each item. The results of the analysis are presented in a descriptive narrative, together with a summary table, to give a clearer picture of the quality of the items studied.

FINDINGS AND DISCUSSION

Item Quality Based on the Content Aspect

An examination of the 40 multiple-choice items in the Final Summative Assessment covering Material Scope or ASALM for Grade X Indonesian language shows that, for the most part, the content aspect already meets the criteria for good items, although the researcher still found some notes on how well the items matched the indicators, matched the learning goals, and how accurate the tested material was.

Indicator Match

Most of the items (32 out of 40, or 80%) already match the learning achievement indicators used as the reference, namely the ability to understand the structure, language features, and writing rules of negotiation texts (items 1-

8, 11-16 and 33), as well as the ability to understand the content and elements of classic and modern literary texts such as *hikayat,* novels, poems, *pantun,* and *gurindam* (items 9, 10, 26–29, and 31). However, the researcher found some items with overlapping indicators — for example, items number 4 and 13, which both test the same indicator, namely the structure of negotiation texts, so that one of the items tends to be redundant and does not add much to the coverage of the indicator being measured.

Learning Goal Match

Looking at the match with the learning goals, the items about negotiation (numbers 1-8, 11-16, 30, and 33, a total of 16 items, or 40% of all items) consistently point to the learning goal of understanding the structure, language rules, and manners of negotiation. This makes the proportion of this material quite dominant compared to literary and biographical material (*hikayat*, novels, poems, *pantun, gurindam,* and biographies), which is represented by 9 items, and discussion and report or review material, represented by 12 items. Because negotiation material makes up 40% of the items, the largest share of any single material, there is a risk that the assessment will not proportionally represent the whole competence that should be measured within that scope of material, considering that *Kurikulum Merdeka* calls for a balance between the competencies of listening, reading-viewing, speaking-presenting, and writing. (Kemendikbudristek, 2022).

Accuracy of Material

In terms of the accuracy of the material, the researcher found some items whose content was less accurate or could confuse students. Item number 27, for example, asks students to interpret an expression in a song lyric excerpt, but the answer options given *("Terlepas," "Berpisah," "Putus hubungan," "Melepas ikatan tali," "Terbebas dari masalah"* - all close in meaning to "released" or "free") are very similar in meaning, so they can easily lead to more than one interpretation. Item number 33 is also less accurate in terms of content, because the answer option "All of the statements above are correct" tends to be too easy to guess, without really testing students' conceptual understanding of how negotiation is applied in context. In addition, item number 31, which asks about the features of a classic story, contains a quoted passage with a typographical error (the word "*klasik*" is repeated in a position that should contain another word, such as "*rakyat*"), which can disrupt students' understanding of the reading context, even though it does not change the intended answer.

Table 1. Distribution of Item Content Based on Learning Scope

No	Scope of Material	Item Number	Total	Percentage
1.	Negotiation Text (concept, structure, attitude, language, dialogue)	1-8, 11-16, 30, 33	16	40%
2.	Classic and Modern Literary Text (*hikayat*, novel, poem, *pantun*, *gurindam*, biography)	9, 10, 24, 26–29, 31, 32	9	22.5%
3.	Discussion Text (response, rebuttal, critical attitude)	17–20, 23, 37	6	15%
4.	Report and Review Text	34–36, 38–40	6	15%
5.	Spelling and Grammar Rules	21, 22, 25	3	7.5%

Overall, the content aspect of all the items is relatively good, because most of the items refer to concrete texts (dialogue excerpts, *hikayat* excerpts, novel excerpts, poem excerpts, and article excerpts), so students are not only asked to remember definitions, but also to understand the context. This matches Arikunto's (2018) view that a good item must measure competence accurately and stay relevant to the competence meant to be measured, not just test memorization of terms.

Item Quality Based on the Construction and Language Aspects

Construction Aspect

In terms of construction, most of the items (around 85%) are written with a clear item stem that does not give a direct clue toward the correct answer. For example, items number 1 to 8 are written as questions or incomplete sentences (fill-in-the-blank sentences) that require students to complete them based on their understanding of the concept, not by guessing from the sentence structure. However, the researcher also found some fairly clear construction weaknesses:

1) Lack of homogeneity of answer options. In item number 13, the answer options are not homogeneous in form: options A and B are sequences of negotiation text structure, while options C, D, and E instead contain structures from other text types (exposition and anecdote). This kind of

unevenness can unintentionally give students a clue that the relevant answer is only in option A or B, which reduces how well the other options work as distractors.

2) Unbalanced length of answer options. Items 8, 11, 33, and 38 have answer options with clearly different sentence lengths — for example, one option is just one word or a short phrase, while another option is a long sentence. This risks becoming an indirect signal for students to avoid, or instead pick, the option that looks "different" from the others, without really using content-based reasoning.

3) Use of "all correct" or "all incorrect" answer options. Item number 33 uses the option "All of the statements above are correct," which, according to the rules for writing multiple-choice items, should be avoided, because it can reduce how effective the distractors are and tends to make it easier for students to answer correctly without full content-based reasoning (Kemendikbudristek, 2022).

Overall, the construction weaknesses found are minor and do not damage the overall validity of the items, but they still need attention from item writers, so that the distractors truly measure understanding, not just serve as a visual difference.

Language Aspect

In terms of language, the researcher found some fairly significant weaknesses, especially related to spelling and punctuation. Items number 21, 22, and 25 explicitly test spelling errors (for example, the capitalization of the words "*sembako*" and "*anda*," and the writing of the month name "*januari*," which should be capitalized). However, in the item stem and text excerpts in other parts (for example, the text in item number 19, which writes "*zat besi*" inconsistently, and item number 9, which writes "*keturuanya*" and "*bisamemberikan*"), the researcher still found unintended typing or spelling errors. This shows an inconsistency between the items that deliberately test spelling accuracy and the quality of language editing in the item texts themselves.

In addition, some item stems still use less effective sentence structures, such as unnecessary word repetition ("*klasik...klasik...klasiknya...klasik*" in item number 31), which could confuse students in understanding the meaning of the sentence. Using clear, communicative language that does not lead to more than

one interpretation is one of the important rules in writing items, as stated by Kemendikbudristek (2022) and Arikunto (2018): item language must be clear, direct, and suited to students' stage of development, so it does not cause misunderstanding of meaning that could affect the validity of the measurement.

Dominance of Cognitive Level in the Items

As shown in Table 2, the researcher analyzed the cognitive level of the 40 items based on the revised Bloom's Taxonomy by Anderson and Krathwohl (2001), which divides the cognitive process into remembering (C1), understanding (C2), applying (C3), analyzing (C4), evaluating (C5), and creating (C6). The analysis shows that the items in the Final Summative Assessment covering Material Scope or ASALM for Grade X Indonesian language are still dominated by low to medium cognitive levels.

Table 2. Distribution of Items Based on Cognitive Level

Cognitive Level	Category	Item Number	Total	Percentage
C1 (Remembering)	LOTS	2, 3, 6, 17, 23, 36, 37	9	22.5%
C2 (Understanding)	LOTS	1, 4, 5, 7, 8, 11, 12, 13, 14, 15, 16, 22, 28	15	37.5%
C3 (Applying)	MOTS	21, 25, 27, 30	4	10%
C4 (Analyzing)	HOTS	9, 10, 19, 20, 26, 29, 31, 32, 35, 39	10	25%
C5 (Evaluating)	HOTS	18, 24, 34, 38, 40	5	5%
C6 (Creating)	HOTS	–	0	0%

The grouping of item numbers in Table 2 is only indicative, because one item can involve more than one cognitive process at the same time — for example, requiring both understanding and light analysis of a text excerpt. The number given refers to the cognitive process that is most prominent in answering that item.

Based on the table above, items in the LOTS category (C1–C2) reach 24 items, or around 60% of all items, while items in the MOTS (Medium Order Thinking Skills) category (C3) are only 4 items (10%), and items in the HOTS category (C4–C6) are 12 items, or around 30%. The researcher did not find a single item that requires the creating (C6) cognitive process, which is naturally hard to achieve in a multiple-choice format, because C6 requires students to

produce a new product or idea, which is better measured through essay tests, projects, or portfolios.

The dominance of LOTS items is clear in items that test definitions and terms (numbers 1, 2, 3, 6, 17, 23, 36, and 37), which in general can be answered just by recalling a concept, without needing to examine the reading context. Meanwhile, HOTS items generally appear in parts that include long text excerpts, such as excerpts from a *hikayat* (number 9), a novel (number 10), health articles (numbers 19 and 20), a poem (number 26), a *gurindam* (number 29), a biography (number 32), reviews (numbers 34 and 35), and observation reports (numbers 38 and 39), which require students to analyze the relationship between parts of the text or infer implied meaning. This pattern supports the finding of Listiani and Rachmawati (2022) that items based on excerpts or authentic stimuli are more likely to measure higher order thinking skills than items that simply ask for a direct definition.

This finding is consistent with the condition identified in the researcher's initial observation, namely that most of the PAT items for Grade X Indonesian language were still dominated by questions that measure lower order thinking skills, while items that require analysis, evaluation, and critical judgment of a text were still limited in number.

Implications of Item Quality for *Kurikulum Merdeka* Assessment

Implications for Assessment Validity

Assessment validity refers to the extent to which an instrument truly measures what it is meant to measure (Sajjad et al., 2020). The dominance of negotiation text material, which makes up more than half of all items, has an impact on the content validity of the assessment, making it less proportional. This is because the scope of Grade X Indonesian language material should cover a more balanced range of competencies among negotiation texts, narrative texts (*hikayat* and novels), poetry, and non-fiction texts (reports, discussions, reviews). This imbalance in material proportion risks making the assessment results less representative of students' overall competence within the tested scope of material, in line with the principle that the important parts of the construct being measured should be proportionally and fairly represented in the final instrument.

On the other hand, construction weaknesses such as the inconsistent homogeneity of answer options (item number 13) and the use of the "all correct" option (item number 33) risk lowering construct validity, because students may answer correctly through a strategy of eliminating options that are not homogeneous, rather than through real mastery of the concept meant to be measured.

Implications for Measuring Critical Thinking Skills

The dominance of LOTS items (60%) over HOTS items (30%) has an impact on limiting the assessment's ability to fully measure students' critical thinking skills. *Kurikulum Merdeka* emphasizes strengthening critical reasoning competence as one dimension of the Pancasila Student Profile (Kemendikbudristek, 2022), so a proportion of HOTS items that is ideally balanced, or even more dominant, becomes crucial in supporting that dimension. The imbalance between LOTS and HOTS in the Final Summative Assessment covering Material Scope or ASALM instrument indicates that the assessment used has not yet fully encouraged students to practice analyzing, evaluating, and synthesizing information from texts, and instead still focuses largely on remembering and understanding concepts at a textual level.

Implications for the Quality of Indonesian Language Learning Evaluation

Overall, the quality of Indonesian language learning evaluation represented by the Final Summative Assessment covering Material Scope or ASALM instrument is fairly good in terms of content and construction, but it still needs improvement in terms of balancing cognitive levels and language accuracy. This finding strengthens previous studies, such as (Riinawati, 2019), who highlighted the importance of a qualitative study of the language aspect in writing items, and Hariati et al. (2026), who emphasized the need for a thorough analysis of item quality, covering both the quantitative and qualitative aspects, so that the evaluation instrument can truly reflect students' competence as a whole and support the right decisions in learning (Arikunto, 2018).

Recommendations for Improving the Items

Examples of Item Revision

Based on the findings on the construction and language aspects, here are some examples of revisions for items identified as having weaknesses.

Each choice represents a permutation of the negotiation structure: (A) request, fulfillment, offer, agreement, purchase, orientation; (B) orientation,

request, fulfillment, offer, agreement, purchase; (C) orientation, offer, request, fulfillment, agreement, purchase; (D) orientation, request, offer, fulfillment, purchase, agreement; (E) request, orientation, fulfillment, offer, agreement, purchase. Students must now truly grasp the right order rather than just be able to tell which alternative is unrelated to the subject.

The revised edition includes a substitute for option E that is more concrete and pertinent, such as "A community organization meeting," so that students still have to weigh each option individually according to their knowledge of how negotiation is applied, rather than selecting a general option that summarizes all the prior choices.

"The citizens of Galuh Pakuan, who have suffered for seventeen years because of the leader Arya Kebonan … now begin to rise again in the hearts of its people. … they hope that Galuh Pakuan can become a safe, peaceful, and prosperous land again …" The meaning is now clear, so students can more easily understand the features of a classic story.

The original text excerpts for item numbers 9 and 19 (prior to revision) contain spelling mistakes. In item 9: "*keturunanya*" and "*keturuanya*" (descendants), and "*bisamemberikan*" (can provide). In item 19: "*kaitanya*" (connection), "*di butuhkan*" (needed), and "*zatbesi*" (iron). These mistakes are incompatible with the items that assess spelling (numbers 21, 22, and 25). Suggested revision: review all of the source text excerpts once more before using them as item stimuli, correcting the words to "*keturunannya*" (descendants), "*dapat memberikan*" (can provide), "*dibutuhkan*" (needed), "*kaitannya*" (connection), and "*zat besi*" (iron), so that consistency is maintained between the items that test spelling and the spelling quality of the excerpts in other items.

The Ideal Form of an Item

The ideal form for a summative Indonesian language assessment item under *Kurikulum Merdeka* should meet several criteria: (1) the item stem is written clearly and does not give a clue toward the correct answer; (2) the answer options are homogeneous, in form, sentence length, and level of difficulty; (3) avoiding the use of "all correct" or "all incorrect" options; (4) including authentic stimuli, such as text excerpts, dialogues, or discourse relevant to students' real-life context, because stimulus-based items have been shown to be more effective at measuring analysis and evaluation skills than items based only on memorization; and (5) arranging the proportion of cognitive levels in a balanced way, with an

ideal combination of around 30% LOTS, 30% MOTS, and 40% HOTS (Hasibuan et al., 2022) or at least a balanced proportion between LOTS and HOTS, so that the assessment measures not only memory but also students' critical reasoning skills, in line with the demands of *Kurikulum Merdeka*.

Recommendations for Teachers

First, teachers are advised to increase the proportion of HOTS items (C4–C6) by using more authentic stimuli, such as literary and non-literary text excerpts that require students to analyze, compare, or evaluate information, instead of simply recognizing definitions. Second, teachers need to review the homogeneity and length of the answer options before the items are used in an assessment, so that each option truly works as a valid distractor. Third, it is recommended that teachers ask colleagues or *Musyawarah Guru Mata Pelajaran (MGMP)* to do a peer review of the item texts before the assessment is given, using a straightforward checklist that addresses effective sentences, the clarity of the item stem, and spelling in accordance with the Indonesian spelling rules. This stage is necessary since the analysis discovered spelling problems in the source excerpts of items 9 and 19 and a duplicated word in the excerpt of item 31, suggesting that language editing was not performed consistently throughout the instrument in the process of editing the language and spelling of the item texts, to make sure the language quality stays consistent across the whole instrument. Fourth, a more careful mapping of material proportion needs to be done in the item blueprint, so that every scope of material taught is proportionally represented in the assessment instrument, making the content validity of the assessment more assured. Fifth, teachers can use the item-writing guidelines and rules published by Kemendikbudristek (2022) as a technical reference in writing or reviewing items, so that the resulting assessment instruments better meet the expected quality standards.

CONCLUSION

Based on the analysis of 40 multiple-choice items in the Final Summative Assessment covering Material Scope or ASALM for Grade X Indonesian language, it can be seen that the items' overall quality is somewhat good according to the content, building, and language standards utilized in this research, but there are still many areas that need to be improved, but there are still several things that need improvement.

In terms of content, most of the items already match the indicators and learning goals that were set. However, the distribution of material is not yet even, since it is dominated by negotiation text material, which reaches 52.5% of all items, so the representation of Indonesian language ability is not yet truly balanced. In terms of item writing, most of the items already follow the rules for writing multiple-choice items, such as a clear item stem that does not give a direct clue toward the correct answer. Even with that, there are still certain shortcomings, such as inconsistent variation in the answer options, imbalanced option length, and the usage of the "all correct" answer option, which might impact how successful the options are at seeming plausible to students who have not yet grasped the content and at distinguishing them from students who do understand it.

In terms of cognitive level, the items mostly fall into the Lower Order Thinking Skills (LOTS) category, at 60%, while Higher Order Thinking Skills (HOTS) items make up only 30%. This condition shows that the assessment tool has not yet fully been able to measure critical thinking skills, which is one of the requirements of *Kurikulum Merdeka*. Because of this, the researcher suggests that item writers increase the number of HOTS items, improve the quality of item writing and the language used, and balance the distribution of material, so that the resulting assessment instrument is more valid, represents various aspects, and is able to measure students' abilities as a whole.

ACKNOWLEDGEMENTS

We would like to acknowledge the valuable contributions of all individuals and institutions who supported this research. We are especially grateful to our colleagues and lecturers for their insightful discussions and constructive feedback throughout the course of this study. We also thank the anonymous reviewers for their thoughtful comments, which greatly improved the quality of this manuscript.

REFERENCES

Aini, N., Islamia N., & Amanati A. (2023). An Analysis of Teachers Questioning Strategy During the Classroom Interaction in EFL Classroom. *Indonesian Journal of Multidisciplinary Educational Research,* 1(2), 137-148. https://doi.org/10.30762/ijomer.v1i2.1516

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan* (Edisi 3). Bumi Aksara.

Hariati, B. S., Hakim, L., Pratiwi, V., Zakia, A., Azizah, R. C. A., Maheswari, A. D., Salsabila, S. E., Aini, Q. Q., & Rahmawati, T. D. (2026). Analisis kualitas butir soal pilihan ganda elemen dokumen berbasis digital menggunakan ANATES. *Jurnal Jendela Pendidikan, 6*(1), 46–58. https://doi.org/10.57008/jjp.v6i01.1956

Hasibuan, F. H., & Wuriyani, E. P. (2022). Analisis soal materi pada buku teks siswa Bahasa Indonesia kelas X semester 1 SMA Negeri 1 Bilah Hulu. *KODE: Jurnal Bahasa, 11*(2), 167–178. https://doi.org/10.24114/kjb.v11i2.36284

Himawan, R., Hermanto, H., Nurgiyantoro, B., Suyono, S., Widyartono, D., Purwanti, A., Yujing, L., & Pogadaev, V. A. (2024). ITEMAN-based evaluation of end-of-semester assessment items: A case study of language test in Indonesian school context. *Indonesian Journal on Learning and Advanced Education (IJOLAE), 6*(3), 387–402. https://doi.org/10.23917/ijolae.v6i3.23254

Kemendikbud. (2020). *Panduan penilaian oleh pendidik dan satuan pendidikan*. Direktorat Jenderal Pendidikan Dasar dan Menengah, Kementerian Pendidikan dan Kebudayaan.

Kemendikbudristek. (2022). *Panduan pembelajaran dan asesmen pada Kurikulum Merdeka*. Badan Standar, Kurikulum, dan Asesmen Pendidikan, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.

Listiani, W., & Rachmawati, R. (2022). Transformasi taksonomi Bloom dalam evaluasi pembelajaran berbasis HOTS. *Jurnal Jendela Pendidikan, 2*(3), 397–402. https://doi.org/10.57008/jjp.v2i03.266

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.

Sajjad, M., Iltaf, S., & Khan, R. A. (2020). Nonfunctional distractor analysis: An indicator for quality of multiple choice questions. *Pakistan Journal of Medical Sciences, 36*(5), 982–986. https://doi.org/10.12669/pjms.36.5.2439

Sudijono, A. (2017). *Pengantar evaluasi pendidikan*. Rajawali Pers.

Syahriandi. (2014). *Analisis kualitas soal Bahasa Indonesia berdasarkan validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan fungsi distraktor. Jurnal Pendidikan Bahasa dan Sastra Indonesia*, 2(1), 45–58.

Syahriandi. (2017). *Analisis kualitatif aspek kebahasaan pada butir soal Bahasa Indonesia. Jurnal Kebahasaan dan Pendidikan*, 5(2), 102–115.

Wafa, A., Hanafi, & Fattah, Z. (2025). Organization of Learning Materials and Evaluation Learning Citizenship Education in College. *Indonesian Journal of Multidisciplinary Educational Research*, 2(2), 207-215. https://doi.org/10.30762/ijomer.v2i2.3753