Beyond Remembering: Assessment and Evaluation of Middle School Mathematics Learning within the Framework of Bloom's Taxonomy

Oktaviana Nirmala Purba[1*], I Gede Ratnaya[2], I Dewa Ayu Made Budhyani[3], I Gusti Lanang Agung Partawa[4]

[1-4]Universitas Pendidikan Ganesha, Singaraja, Indonesia

Corresponding Author's e-mail: *oktaviana.nirmala@student.undiksha.ac.id*

ARMADA
JURNAL PENELITIAN MULTIDISIPLIN

e-ISSN: 2964-2981

ARMADA : Jurnal Penelitian Multidisiplin

https://ejournal.45mataram.ac.id/index.php/armada

Vol. 04, No. 07 Juli, 2026

Page: 3224-3234

DOI:

https://doi.org/10.55681/armada.v4i7.3203

Article History:

Received: Mei 13, 2026

Revised: Juni 29, 2026

Accepted: Juli 16, 2026

Abstract *: Mathematics assessment in junior secondary schools should respond to students' diverse cognitive attainment across Bloom's Taxonomy, from C1 to C6. However, studies indicate that assessment instruments in textbooks, exercises, daily tests, and end-of-semester examinations remain dominated by C1–C3 questions and inadequately support higher-order thinking skills. This article examines the alignment of mathematics assessment practices with the responsiveness principles of Bloom's Taxonomy. The study employed a qualitative approach using secondary data and a critical qualitative review design. Data came from articles, research reports, and review documents indexed in Google Scholar, ERIC, and arXiv, and were analysed thematically and critically. The findings reveal a gap between HOTS-oriented curriculum expectations and the cognitive distribution of questions. In one Year 7 mathematics textbook, C2 questions accounted for 75.2%, whereas C6 questions represented only 0.4%. These findings highlight the need for more balanced, adaptive, and responsive mathematics assessments that accommodate variations in students' thinking abilities.*

Keywords : *Assessment; Evaluation Of Mathematics Learning; Bloom's Taxonomy; Junior High School; Critical Review*

Abstrak : Asesmen matematika di Sekolah Menengah Pertama (SMP) idealnya responsif terhadap keberagaman capaian kognitif peserta didik berdasarkan Taksonomi Bloom dari C1 hingga C6. Namun, berbagai kajian menunjukkan bahwa instrumen asesmen pada buku teks, latihan, ulangan harian, dan Penilaian Akhir Semester masih didominasi soal level C1–C3, sehingga belum mendukung penguatan kemampuan berpikir tingkat tinggi. Artikel ini bertujuan mengkaji kesesuaian praktik asesmen dan evaluasi matematika di SMP dengan prinsip responsivitas Taksonomi Bloom. Penelitian menggunakan pendekatan kualitatif berbasis data sekunder dengan desain tinjauan kualitatif kritis. Data diperoleh dari artikel ilmiah, laporan penelitian, dan dokumen kajian melalui Google Scholar, ERIC, dan arXiv, kemudian dianalisis secara tematik-kritis. Hasil telaah menunjukkan kesenjangan antara tuntutan kurikulum berbasis HOTS dan distribusi kognitif soal. Pada salah satu buku teks matematika kelas VII, soal C2 mencapai 75,2%, sedangkan soal C6 hanya 0,4%. Temuan ini menegaskan perlunya pengembangan asesmen matematika yang lebih seimbang, adaptif, dan responsif terhadap variasi kemampuan berpikir peserta didik.

Kata Kunci : Asesmen; Evaluasi Pembelajaran Matematika; Taksonomi Bloom; Sekolah Menengah Pertama; Tinjauan Kritis

INTRODUCTION

The results of the 2022 Programme for International Student Assessment (PISA) constitute a serious warning for the Indonesian mathematics education sector. Indonesia achieved an average mathematics score of only 366, while the 2022 assessment placed particular emphasis on mathematical reasoning and students' capacity to apply mathematics in real-life situations (OECD, 2023). Moreover, only approximately 18% of Indonesian students aged 15 attained at least the minimum proficiency Level 2 in mathematics, far below the OECD average of 69%. This indicates that more than eight out of ten Indonesian students remain unable to interpret and represent simple situations mathematically without direct instruction (OECD, 2023). Such findings are not merely international statistical indicators detached from classroom realities; they reflect the extent to which mathematical reasoning is developed and assessed in schools. Mathematical competence includes not only procedural mastery but also reasoning, modelling, problem-solving, representation, and communication, which must be cultivated systematically throughout mathematics education (Niss & Højgaard, 2019). This issue is particularly important at the junior secondary school level, which represents a crucial period for developing mathematical reasoning before students progress to more advanced education.

Ironically, despite these low levels of achievement, mathematics assessment practices in Indonesian junior secondary classrooms remain concentrated at relatively basic cognitive levels. An analysis of an official Year 7 mathematics textbook revealed that, among 537 questions, 75.2% were classified at the understanding level (C2), whereas questions at the analysing (C4), evaluating (C5), and creating (C6) levels accounted for only 3.9%, 0.2%, and 0.4%, respectively. No questions were identified at the remembering level (C1) (Anifarka & Rosnawati, 2023). A similar pattern was identified in exercises and competency-test questions concerning Systems of Linear Equations in Two Variables in a Year 8 mathematics textbook. Most questions were limited to understanding (C2) and applying (C3), while none required evaluating (C5) or creating (C6) (Rikawati, 2021). The dominance of lower cognitive demands was also found in the analysis of Year 9 end-of-semester mathematics assessments, in which lower-order questions substantially outnumbered higher-order thinking questions (Azizah et al., 2021).

These findings are consistent with broader research on mathematics textbooks. Wijaya et al. (2015) found that only approximately 10% of tasks in the Indonesian mathematics textbooks they analysed were context-based. Of those tasks, only 2% required reflection, which represented the highest level of cognitive demand in their analytical framework. More recent comparative research has likewise shown that the proportion of analysing, evaluating, and creating tasks differs significantly across countries and textbook systems, suggesting that textbooks can either facilitate or restrict students' opportunities to develop HOTS (Chawla et al., 2026). Thus, Indonesian junior secondary students are more frequently assessed through routine understanding and procedural application than through reasoning, evaluation, reflection, and the creation of mathematical strategies. This pattern corresponds closely with Indonesia's low PISA mathematics achievement because PISA primarily assesses mathematical literacy and reasoning rather than the mechanical reproduction of formulas.

Theoretically, Bloom's Taxonomy was initially formulated by Bloom and his colleagues in 1956 as a framework for classifying educational objectives. It was subsequently revised into a two-dimensional framework encompassing the knowledge dimension and six cognitive-process categories: remembering (C1), understanding (C2), applying (C3), analysing (C4), evaluating (C5), and creating (C6) (Krathwohl, 2002; Listiani & Rachmawati, 2022). In mathematics education, the taxonomy is important because the hierarchical and interconnected nature of mathematical knowledge requires students not only to memorise formulas and understand concepts but also to apply procedures, analyse patterns, evaluate solution strategies, and construct original mathematical models or solutions (Effendi, 2017). The first three cognitive categories are

commonly associated with Lower-Order Thinking Skills (LOTS), whereas analysing, evaluating, and creating are generally associated with Higher-Order Thinking Skills (HOTS).

Nevertheless, HOTS should not be interpreted merely as a collection of difficult questions. It encompasses problem-solving, critical thinking, metacognition, teamwork, and innovative or creative development (Liu et al., 2024). Consequently, the cognitive category of an assessment item should be determined by the mental processes required to solve it rather than by its perceived level of difficulty. Research has demonstrated that assessment items can measure both lower- and higher-order cognitive processes, but their design, stimulus materials, and scoring procedures must be carefully aligned with the intended cognitive construct (Haataja et al., 2023). The operational verbs associated with each level of Bloom's Taxonomy therefore play an important role in guiding teachers to formulate measurable learning objectives and assessment questions, although verbs alone cannot determine cognitive complexity without considering the content, context, and reasoning demanded by the task (Newton et al., 2020).

The principle of constructive alignment requires intended learning outcomes, learning activities, and assessment instruments to be directed towards corresponding cognitive processes (Biggs, 1996). An assessment may therefore be considered responsive to Bloom's Taxonomy when its items provide balanced opportunities for students to demonstrate diverse cognitive achievements rather than concentrating almost exclusively on one or two levels. Empirical alignment studies have shown that discrepancies can occur among curriculum standards, learning objectives, classroom activities, and assessment tasks when teachers do not interpret or implement cognitive demands consistently (Zhao et al., 2023). In mathematics education, such discrepancies may cause students to experience learning activities at one cognitive level while being assessed at another or, conversely, to encounter curriculum objectives that emphasise reasoning while classroom assessments remain procedural.

For example, a learning objective stating that "students are able to solve systems of linear equations in two variables" may appropriately be measured through a C3 task requiring students to apply elimination or substitution procedures. However, when the objective states that "students are able to evaluate the efficiency of different methods for solving systems of linear equations in two variables and create alternative strategies for particular cases", the corresponding assessment must require C5–C6 processes rather than routine calculation. The use of verbs such as "analyse" or "evaluate" in curriculum documents does not automatically produce HOTS-oriented learning when the assessment task only requires the reproduction of a familiar procedure. The discrepancy between ambitious HOTS-oriented learning objectives and questions that remain at C2–C3 constitutes a tangible form of curricular misalignment.

The gap between HOTS-oriented curriculum policies and teachers' competence in constructing cognitively responsive mathematics assessments is not merely a technical issue involving question formulation. It also reflects structural challenges related to teachers' assessment literacy, professional development, curriculum interpretation, and access to examples of high-quality assessment tasks. A qualitative study conducted in Nigeria found that science and mathematics teachers tended to formulate objectives and assessment questions at basic cognitive levels. Analysis, evaluation, and creation were addressed less frequently because teachers lacked adequate pre-service and in-service preparation concerning the appropriate application of Bloom's Taxonomy (Muhammad & Rashid, 2024). In addition, teachers' judgements of students' HOTS performance may be influenced by students' preceding LOTS performance, creating potential bias when different cognitive abilities are not assessed and interpreted separately (Jansen & Möller, 2022). The similarities between the findings reported in Nigeria and Indonesia suggest that unequal cognitive distributions in assessment represent a broader structural challenge in education systems, particularly in mathematics and science, where reasoning develops progressively and hierarchically.

Assessment should also function as an integral component of learning rather than solely as a mechanism for assigning scores. Formative assessment becomes effective when evidence of students' understanding is used to adapt instruction, provide feedback, and help students progress towards intended learning goals (Black & Wiliam, 2009). Therefore, a responsive mathematics assessment should provide information not only about whether students can produce a correct

answer but also about how they interpret problems, select strategies, justify procedures, evaluate alternative solutions, and construct mathematical arguments. An instrument dominated by C2–C3 questions provides limited evidence concerning these more complex dimensions of mathematical thinking.

If mathematics assessments in junior secondary classrooms continuously stop at remembering, understanding, and routine application, it is difficult to expect students to demonstrate advanced mathematical reasoning when participating in national or international assessments such as PISA. The discrepancy between the mathematics taught and tested in classrooms and the mathematics demanded by the curriculum and contemporary society therefore requires critical examination. This concern has become increasingly relevant under the direction of the Kurikulum Merdeka, which promotes diagnostic assessment and deep learning, as well as the implementation of the Minimum Competency Assessment or Asesmen Kompetensi Minimum (AKM), which emphasises contextual literacy and numeracy rather than procedural mastery alone.

Numeracy within the AKM framework requires students to apply mathematical knowledge and reasoning in personal, social, and scientific contexts. Conceptually, these competencies correspond strongly with analysing, evaluating, and creating processes in Bloom's Taxonomy. However, when daily assessments and textbook questions remain dominated by C2–C3 tasks, the discrepancy between the principles of AKM and Kurikulum Merdeka and actual classroom practice is likely to persist. This concern is supported by findings indicating that contextual numeracy content and higher-order cognitive demands remain limited in junior secondary mathematics textbooks (Anifarka & Rosnawati, 2023; Wijaya et al., 2015).

Based on these empirical problems and theoretical considerations, this article aims to critically examine how mathematics assessment and evaluation practices at the junior secondary school level are represented in scientific literature and previous research reports. It specifically investigates the extent to which these practices correspond with the responsiveness principles of Bloom's Taxonomy and identifies the implicit assumptions, structural disparities, and assessment biases that may restrict the recognition and development of students' diverse thinking abilities. The investigation is conducted using a qualitative approach based on secondary data and a critical qualitative review design.

RESEARCH METHODS

This study employed a qualitative approach based on secondary data using a critical qualitative review design. The research was conducted through four iterative stages: determining the focus and critical questions, searching and screening sources, conducting thematic mapping and critical analysis, and synthesising the findings into an argumentative framework concerning the responsiveness of Bloom's Taxonomy in junior secondary mathematics assessment. The review addressed three questions: the distribution of cognitive levels in assessment instruments, the assumptions underlying disparities in that distribution, and their implications for accommodating students' diverse thinking abilities. The unit of analysis comprised scientific documents rather than human participants. Sources were selected purposively from Google Scholar, ERIC, and arXiv and were limited to indexed articles or institutional reports published between 2016 and 2026 that discussed mathematics assessment in relation to Bloom's cognitive levels. Sources lacking clear methods, containing unsupported opinions, or duplicating previous findings were excluded. Four studies were analysed most intensively because they provided explicit data on cognitive-level distribution and test-item quality (Anifarka & Rosnawati, 2023; Rikawati, 2021; Azizah et al., 2021; Kurniasi et al., 2020).

The research materials included journal articles, policy reports, item-analysis studies, and previous research documents. Data were organised using a literature-search matrix, thematic coding sheet, and analytical memos. The thematic instrument recorded the source, year, educational level, method, distribution of C1–C6 levels, and critical notes. It was piloted on several sources to ensure consistent classification despite differences in terminology, such as "HOTS questions" and "C4–C6 questions". Data analysis involved thematic mapping, critical analysis of gaps between curriculum requirements and assessment practices, and critical synthesis to identify patterns and formulate recommendations for more responsive assessment. Iterative reading and

reflective memo-writing were used throughout the process to maintain credibility and minimise confirmation and interpretation bias.

RESULTS AND DISCUSSION

Results The thematic mapping of the secondary sources studied shows a consistent pattern in junior high school mathematics learning: assessment instruments, whether in the form of example questions and exercises in textbooks, chapter competency test questions, daily quizzes, or End-of-Semester Assessments, are generally still dominated by low to medium cognitive levels (C1–C3), while high cognitive levels (C4–C6) are present in much smaller proportions, and in some cases, are nearly nonexistent. The results are presented in three sub-themes: (1) distribution of cognitive levels in mathematics textbooks; (2) distribution of cognitive levels in practice questions versus competency test questions for specific materials; and (3) cross-grade comparison for grades VII–IX.

Cognitive Level Distribution In Seventh Grade Mathematics Textbooks

The analysis by Anifarka and Rosnawati (2023) of 537 items in the official first semester seventh-grade junior high school mathematics textbook, which includes 88 example questions and 447 "Let's Practice" and "Competency Test" questions, reveals a highly uneven distribution of cognitive levels, as presented in Table 1.

Table 1. Distribution of Cognitive Levels in Seventh Grade Junior High School Mathematics Textbooks (n=537)

Level Kognitif	Deskripsi	Number of Questions	Percentage
C1 – Remembering	Reiterating terms/formulas	0	0%
C2 – Understanding	Explaining the concept in your own words	≈ 404	75,2%
C3 – Applying	Using formulas in routine situations	≈ 109	20,3%
C4 – Analyzing	Sorting elements/patterns in the problem	≈ 21	3,9%
C5 – Evaluating	Evaluating the correctness of a solution	≈ 1	0,2%
C6 – Creating	Developing an original model/solution	≈ 2	0,4%

Source: Processed from Anifarka & Rosnawati (2023)

The data in Table 1 shows that more than three-quarters of the questions (75.2%) are at the understanding level (C2), while the three HOTS levels (C4, C5, and C6) combined only account for about 4.5% of the total questions. In fact, the remembering level (C1) is completely absent, indicating that the textbook is indeed designed to directly emphasize conceptual understanding, but does not proportionally progress to the levels of analysis, evaluation, and creation. The same study also noted that numeracy questions with personal, social, or scientific contexts, as emphasized in the Minimum Competency Assessment (AKM), are still rarely included in textbook questions.

This finding is important to note considering the position of textbooks as the main reference for teachers in preparing teaching materials and other derivative questions. When 75.2% of the questions in the official textbook stop at the understanding level, teachers who use the textbook as their main reference have a high potential to replicate a similar pattern, consciously or unconsciously, in the daily quizzes and Semester Final Assessments they prepare themselves. In other words, the disparity in Table 1 is not just an issue with one textbook, but rather has the potential to become a cognitive "template" that is repeatedly reproduced throughout the academic year in various forms of classroom assessments.

Distribution Of Cognitive Levels In Practice Questions And Competency Tests On Spldv Material For Grade Viii

A similar, even more extreme pattern was found in Rikawati's (2021) analysis of practice questions and competency test questions on the topic of Systems of Two Variable Linear Equations

(SPLDV) in the 8th-grade first semester mathematics textbook published by the Ministry of Education and Culture. The comparison of the distribution of the two types of questions is presented in Table 2.

Table 2. Distribution of Cognitive Levels of Practice Questions vs Competency Test Questions on SPLDV Material for Grade VIII

Level Kognitif	Soal Latihan (n=40)	Soal Uji Kompetensi (n=20)
C1 – Remembering	0 (0%)	0 (0%)
C2 – Understanding	9 (22,5%)	2 (10%)
C3 – Applying	26 (65%)	18 (90%)
C4 – Analyzing	5 (12,5%)	0 (0%)
C5 – Evaluating	0 (0%)	0 (0%)
C6 – Creating	0 (0%)	0 (0%)

Source: Processed from Rikawati (2021)

Table 2 shows a striking pattern: in the competency test questions, which are supposed to be the "peak" of material mastery evaluation, 90% of the questions are at the applying level (C3), with an even narrower distribution compared to regular practice questions. Not a single one of the 60 analyzed questions (both practice and competency tests) touches the evaluating level (C5) or creating level (C6). These findings indicate that the more "formal" the position of an assessment instrument in a textbook, the narrower the variation in its cognitive levels, a pattern that contradicts the assumption that competency tests should demand a more comprehensive and tiered mastery.

The phenomenon of the narrowing of cognitive levels in competency test questions compared to practice questions (Table 2) can also be interpreted as an indication of an assessment orientation that prioritizes the practicality of scoring over the depth of reasoning. Questions at the C3 level (applying the SPLDV formula directly) are much easier to assess objectively and quickly compared to C5 or C6 level questions that require evaluation of the quality of arguments or originality of students' solutions. This tendency is consistent with the pattern in Table 1 and reinforces the suspicion that textbook authors and teachers tend to prioritize the administrative efficiency of assessments over the depth of cognitive achievements that are intended to be measured.

Comparison Across Grades Vii–Ix

To obtain a more comprehensive picture, the findings from the four sources covering grades VII to IX are summarized comparatively in Table 3.

Table 3. Comparison of Findings on the Distribution of Cognitive Levels in Cross-Grade Middle School Mathematics Assessment

Level/Class	Source (Year)	Research Object	Main findings
VII	Anifarka & Rosnawati (2023)	Research Object Mathematics textbook (537 questions)	C2 dominates (75.2%); combined C4–C6 only 4.5%
VIII	Rikawati (2021)	Practice questions & SPLDV competency test (60 questions)	C3 dominates (65–90%); no C5–C6 questions
IX	Azizah et al. (2021)	Mathematics End-of-Semester Assessment questions	Dominated by C1–C3; low proportion of C4–C6
IX	Kurniasi et al. (2020)	Mathematics Daily Test questions (essay)	Focus on conceptual understanding & problem-solving; question quality varies from easy to moderate

Source: Processed from Anifarka & Rosnawati (2023), Rikawati (2021), Azizah et al. (2021), and Kurniasi et al. (2020)

Table 3 emphasizes that the gap in responsiveness of Bloom's Taxonomy in middle school mathematics assessments is consistent across grade levels, not merely an issue of a single textbook or a single school. The shift in dominance from C2 (grade VII) to C3 (grades VIII and IX) indicates an increase in cognitive levels as the grade level progresses, but this increase stops at the applying level (C3) and almost never reaches the analyzing level and above (C4–C6) proportionally. The study by Kurniasi et al. (2020) on daily test questions for ninth grade further reinforces this picture from a different perspective: although essay questions have the potential to involve more complex reasoning, the focus of the evaluation remains on understanding concepts and solving routine problems, with the difficulty level of the questions varying from easy to moderate without touching the difficult category, which usually indicates a demand for higher-level reasoning.

If the three tables are read together, a cyclical pattern emerges: students enter junior high school with assessments emphasizing conceptual understanding (C2) in seventh grade, then shift to procedural application (C3) in eighth and ninth grades, but are never truly directed to practice analyzing, evaluating, and creating proportionally throughout the three years of junior high school. Ideally, these three levels of HOTS should have been gradually introduced since seventh grade so that students have a sufficiently mature reasoning foundation when they enter ninth grade and prepare for more demanding reasoning assessments, such as the Final Summative Assessment or later PISA.

This cyclical pattern can also be interpreted as a form of "curricular repetition" at the low to medium cognitive level: instead of forming a curve of gradual but consistent increase in thinking complexity from grade VII to grade IX, the data in Table 3 actually shows a relatively flat curve in the C2–C3 range at all three levels. The implication is that middle school students in Indonesia may complete three years of mathematics education without ever being systematically trained to analyze patterns, evaluate the correctness of a strategy, or create original solutions to mathematical problems. three skills that are actually the main determinants of success in reasoning assessments like PISA and the demands of the 21st-century workforce, which emphasize complex problem-solving and creative thinking.

The Gap with the Spirit of Numeracy in the Minimum Competency Assessment (AKM)

In addition to the distribution of cognitive levels, the findings of Anifarka and Rosnawati (2023) also highlight another equally important dimension: the involvement of numeracy contexts as emphasized in AKM. Out of the 537 questions analyzed, only a small portion involved personal, social, or scientific contexts three contexts that are characteristic of AKM numeracy questions and simultaneously represent real-world mathematical reasoning. The majority of the questions are still in the form of pure mathematics, requiring symbol manipulation without framing them in contextual situations. This gap reinforces the notion that junior high school mathematics textbooks have not fully shifted direction to align with the spirit of the AKM policy and the Merdeka Curriculum, even tho both policies have been implemented nationally.

This finding is important to be read alongside PISA data in the Introduction: if contextual numeracy questions, which are essentially HOTS questions wrapped in real contexts, are rarely practiced in class thru textbooks or daily assessments, it is understandable that Indonesian students struggle with PISA questions, which are almost entirely presented in situational contexts. In other words, the gap in cognitive levels (Table 1–Table 3) and the gap in numeracy contexts intertwine and reinforce each other in explaining the low mathematical achievements of Indonesian junior high school students in large-scale assessments.

Kesenjangan konteks numerasi ini juga berimplikasi pada validitas konstruk (construct validity) instrumen asesmen matematika di sekolah. Apabila tujuan pembelajaran dan kurikulum secara eksplisit menuntut kemampuan numerasi kontekstual sebagaimana digariskan AKM, namun instrumen asesmen yang digunakan guru sehari-hari (baik dari buku teks maupun buatan sendiri) masih didominasi soal matematis murni tanpa konteks, maka terdapat celah antara apa yang hendak diukur (construct yang dituju) dan apa yang benar-benar diukur (construct yang teramati). Celah semacam ini, dalam kerangka constructive alignment yang telah dibahas pada Pendahuluan, merupakan bentuk lain dari ketidakselarasan antara tujuan pembelajaran, aktivitas pembelajaran, dan asesmen kali ini bukan pada dimensi level kognitif semata, melainkan pada dimensi konteks penerapan pengetahuan matematis itu sendiri.

Discussion

Logically, the findings in Tables 1 to 3 can be interpreted as an indication that the gap in the responsiveness of mathematics assessments to Bloom's Taxonomy is layered and systemic. At the macro level, the dominance of LOTS over HOTS occurs consistently from seventh-grade textbooks (Anifarka & Rosnawati, 2023) to eighth-grade competency test questions (Rikawati, 2021) and ninth-grade summative assessments (Azizah et al., 2021). At the micro level, even among questions that are formally positioned as "peak" evaluation instruments such as chapter competency tests, the proportion of HOTS actually decreases compared to regular practice questions (Table 2). This layered pattern aligns with the explanation by Muhammad and Rashid (2024) that teachers tend to remain at the cognitive level that is easiest to formulate and assess, due to the lack of pre-service and in-service training on the proper application of Bloom's Taxonomy in exact subjects like mathematics.

This interpretation is also relevant in relation to the principle of constructive alignment and the list of operational verbs in Bloom's Taxonomy compiled by Newton et al. (2020): the inequality of cognitive distribution is most likely rooted in the difficulties of textbook writers and teachers in formulating indicators and operational verbs for the levels of evaluating (C5) and creating (C6) in the context of mathematics, for example, requiring students to assess the correctness of a solution strategy (C5) or constructing an original mathematical model for a contextual problem (C6) that is much more complex than simply applying a formula to a routine problem (C3). As emphasized by Effendi (2017), the hierarchical characteristics of mathematics should actually demand a balanced combination of procedural exercises (C1–C3) and higher-order reasoning (C4–C6), but the practice of constructing questions in the field seems to prioritize the efficiency of procedural assessments that are easier to score objectively.

A further implication of these findings is the risk of continuity between classroom mathematics assessment and broader-scale assessment. If junior high school mathematics assessment instruments systematically rarely train students to the levels of analysis, evaluation, and creation, then low achievement in international assessments such as PISA mathematics (OECD, 2023) is not simply a matter of curriculum or macro policies, but also the accumulation of repeated micro practices in the classroom from grades VII to IX. Therefore, strengthening the responsiveness of mathematics assessments to Bloom's Taxonomy in junior high schools needs to target three goals simultaneously: (1) revising the proportion of cognitive levels in official textbooks so that they do not skewed towards C2; (2) strengthening the variety of cognitive levels in competency test questions and summative assessments, not just practice questions; and (3) increasing teacher capacity in formulating indicators and operational verbs for levels C4–C6 concretely in the context of junior high school mathematics material.

As a critical note on the sources reviewed, it is important to acknowledge that the four main studies (Anifarka & Rosnawati, 2023; Rikawati, 2021; Azizah et al., 2021; Kurniasi et al., 2020) have limited scope: each analyzed only one textbook, one material, or one specific school, so generalizations across regions and across curricula require caution. However, the consistent direction of the findings, with the dominance of C2–C3 and the scarcity of C5–C6 across various grade levels, regions, and types of instruments (textbooks, practice questions, daily tests, and PAS), actually strengthens the external validity of the patterns found, rather than weakening them. A critical reading of the small variations in the methodologies of each study (quantitative descriptive, qualitative descriptive, and document analysis) also demonstrates that the findings of cognitive level disparities are robust to differences in research approaches, further strengthening the belief that this issue is structural, not simply a methodological artifact of a single study.

From a policy perspective, these findings imply that interventions that focus solely on training teachers to develop "HOTS" questions without comprehensive revisions to textbooks and summative assessment guidelines risk missing the root of the problem. Textbooks, as an "experienced curriculum," strongly influence teachers' mindsets when developing derived questions. If the textbook's sample questions and exercises skew toward C2–C3, teachers tend to replicate the same patterns in their own daily tests and self-assessed PAS. Therefore, Bloom's Taxonomy-responsive mathematics assessment reform should ideally begin at the top, namely

textbooks and official question banks, before requiring changes in practices at the individual teacher level.

In closing, the findings above confirm the main thesis of this article: that junior high school mathematics assessment in Indonesia is still systematically "stopping at remembering and understanding" and has not "transcended" that level to the reasoning, evaluation, and creation that characterize Higher Order Thinking Skills. As long as this shift has not occurred comprehensively, from textbooks, practice questions, competency tests, to summative assessments, the gap between junior high school students' achievements in the classroom and their achievements on national and international assessments such as PISA will likely continue to recur in subsequent assessment cycles.

CONCLUSIONS AND RECOMMENDATIONS

This critical qualitative review based on secondary data shows that assessment and evaluation practices for mathematics learning in junior high schools still face multiple and consistent gaps across levels regarding the responsiveness principle of Bloom's Taxonomy. In the seventh-grade textbook, the proportion of questions at the understanding level (C2) reaches 75.2%, while the three HOTS levels combined are only around 4.5%. In the eighth-grade SPLDV material, even the competency test questions, which should be the culmination of the evaluation, are 90% at the applying level (C3), with not a single question at the evaluating or creating level. A similar pattern continues in the ninth-grade summative assessment and daily tests. This multiple gap is consistent with the low achievement of Indonesian students on international assessments such as the PISA mathematics test and indicates that the root of the problem lies not solely in curriculum policy but also in the limited capacity of teachers and textbook authors to translate HOTS demands into mathematics assessment instruments that are proportionately cognitively tiered.

Further research is recommended to combine this critical review with primary data in the form of document analysis of mathematics assessment questions and teacher interviews at selected junior high schools to examine the extent to which the patterns of inequality identified in the literature are actually manifested in practice, and to expand the coverage of other mathematics materials. For textbook developers and curriculum policymakers, it is recommended that the proportion of cognitive levels in example questions, practice questions, and competency test questions be reviewed for greater balance. For school administrators and teacher professional development providers, it is recommended that training emphasize skills in formulating mathematics assessments at higher cognitive levels in a hierarchical manner, particularly the less frequently addressed levels of evaluation and creation, rather than simply meeting HOTS question quantity targets.

As a critical qualitative review based on secondary data, this study has limitations that need to be openly acknowledged: first, the secondary sources reviewed in depth were limited to a few textbooks and specific schools, thus not representing the full range of mathematics textbooks and assessment practices across Indonesia; second, this study did not involve primary data collection, so the critical interpretation presented is a synthesis of the findings of others, not a direct verification in the field; and third, the publication period of the sources reviewed (2016–2026) allows for recent policy changes or textbook revisions that were not fully captured in the literature available at the time of this study. These limitations open up opportunities for further confirmatory research using primary data.

ACKNOWLEDGMENTS

The authors express their sincere gratitude to Ganesha University of Education for all the academic support that facilitated the completion of this research. The authors also appreciate the encouragement, insights, and valuable constructive feedback from colleagues throughout the research and writing process. In addition, appreciation is also expressed to all parties and institutions that have provided moral and material support, which contributed significantly to the successful completion of this research.

DAFTAR PUSTAKA

Anifarka, A., & Rosnawati, R. (2023). Analisis buku teks matematika SMP berdasarkan tingkat kognitif pada Taksonomi Bloom revisi dan numerasi pada AKM. Jurnal Cendekia: Jurnal Pendidikan Matematika, 7(3), 2151–2166. https://doi.org/10.31004/cendekia.v7i3.1701

Azizah, N., Hanifah, & Sumardi, H. (2021). Analisis kualitas dan tingkat kognitif soal matematika Penilaian Akhir Semester (PAS) ganjil kelas IX di SMPN 10 Kota Bengkulu tahun 2020/2021. Journal Mathematics Education Sigma (JMES), 2(2). https://doi.org/10.30596/jmes.v2i2.7936

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364. https://doi.org/10.1007/BF00138871

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Chawla, K., Phatak, J. U., & Chawla, R. (2026). Comparison of high school textbooks analysing higher-order thinking skills development in India and Singapore. *Thinking Skills and Creativity, 60*, 102058. https://doi.org/10.1016/j.tsc.2025.102058

Effendi, R. (2017). Konsep revisi Taksonomi Bloom dan implementasinya pada pelajaran matematika SMP. JIPMat: Jurnal Ilmiah Pendidikan Matematika, 2(1), 72–78. https://doi.org/10.26877/jipmat.v2i1.1483

Haataja, E. S. H., Tolvanen, A., Vilppu, H., Kallio, M., Peltonen, J., & Metsäpelto, R.-L. (2023). Measuring higher-order cognitive skills with multiple-choice questions: Potentials and pitfalls of Finnish teacher education entrance. *Teaching and Teacher Education, 122*, 103943. https://doi.org/10.1016/j.tate.2022.103943

Jansen, T., & Möller, J. (2022). Teacher judgements in school exams: Influences of students' lower-order-thinking skills on the assessment of students' higher-order-thinking skills. *Teaching and Teacher Education, 111*, 103616. https://doi.org/10.1016/j.tate.2021.103616

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218. https://doi.org/10.1207/S15430421TIP4104_2

Kurniasi, E. R., Yopa, Y., & Karennisa, F. (2020). Analisis soal ulangan harian matematika kelas IX SMP Negeri 1 Toboali. Jurnal Ilmu Pendidikan (JIP) STKIP Kusuma Negara, 12(1), 43–52. https://doi.org/10.37640/jip.v12i1.276

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

Liu, J., Liu, Z., Wang, C., Xu, Y., Chen, J., & Cheng, Y. (2024). K–12 students' higher-order thinking skills: Conceptualisation, components, and evaluation indicators. *Thinking Skills and Creativity, 52*, 101551. https://doi.org/10.1016/j.tsc.2024.101551

Muhammad, U., & Rashid, R. A. A. (2024). Teachers' application of Bloom's taxonomy in the teaching and assessment of critical thinking skills in Nigeria. International Journal of Education, Psychology and Counselling, 9(55), 126–137. https://doi.org/10.35631/IJEPC.955007

Newton, P. M., Da Silva, A., & Peters, L. G. (2020). A pragmatic master list of action verbs for Bloom's taxonomy. Frontiers in Education, 5, 107. https://doi.org/10.3389/feduc.2020.00107

Niss, M., & Højgaard, T. (2019). Mathematical competencies revisited. *Educational Studies in Mathematics, 102*(1), 9–28. https://doi.org/10.1007/s10649-019-09903-9

OECD. (2023). PISA 2022 results (Volume II): Learning during – and from – disruption. OECD Publishing. https://doi.org/10.1787/a97db61c-en

Rikawati, N. (2021). Analisis tingkat kognitif soal matematika kelas VIII berdasarkan Taksonomi Bloom revisi. Journal Mathematics Education Sigma (JMES), 2(2), 61–71. https://doi.org/10.30596/jmes.v2i2.7939

Wijaya, A., van den Heuvel-Panhuizen, M., & Doorman, M. (2015). Opportunity-to-learn context-based tasks provided by mathematics textbooks. *Educational Studies in Mathematics, 89*(1), 41–65. https://doi.org/10.1007/s10649-015-9595-1

Zhao, L., Zhao, B., & Li, C. (2023). Alignment analysis of teaching–learning–assessment within the classroom: How teachers implement project-based learning under the curriculum standards. *Disciplinary and Interdisciplinary Science Education Research, 5*, Article 13. https://doi.org/10.1186/s43031-023-00078-1