

Validity and Reliability Analysis of a Contextual Problem-Based Essay Test For Mathematical Critical Thinking on Cubes and Rectangular Prisms

Imelda^{1*}, Izwita Dewi², KMS Muhammad Amin Fauzi²

¹Department of Mathematics Education, Universitas Katolik Santo Thomas, Medan, Indonesia

²Department of Mathematics Education, Universitas Negeri Medan, Medan, Indonesia

imelda@ust.ac.id, izwitadewi@unimed.ac.id, aminfauzi@unimed.ac.id

corresponding author: imelda@ust.ac.id

ABSTRACT

This study aimed to develop and evaluate a contextual problem-based essay test for assessing students' mathematical critical thinking skills on the topics of cubes and rectangular prisms. A quantitative descriptive design was employed involving 30 ninth-grade students from SMPN 31 Medan during the pilot-testing stage. Prior to implementation, the instrument was evaluated by expert validators to establish content validity. Empirical item validity was analyzed using the Pearson Product-Moment correlation coefficient, while internal consistency reliability was estimated using Cronbach's Alpha. The findings showed that all essay items met the criterion of empirical validity and that the instrument demonstrated moderate internal consistency, indicating its suitability for preliminary classroom assessment and pilot testing. The developed instrument assesses higher-order cognitive processes based on Facione's critical thinking framework, including interpretation, analysis, evaluation, inference, and explanation, through authentic contextual geometry problems. This study contributes to mathematics education by providing a contextually relevant assessment instrument with preliminary psychometric evidence that can support teachers in implementing authentic assessment of mathematical critical thinking skills. The findings also provide a foundation for future large-scale validation studies aimed at developing standardized assessment instruments for mathematics education.

Keywords: *instrument development, mathematical critical thinking assessment, geometry, essay test*

Submitted	Accepted	Published
27 June 2026	19 July 2026	30 July 2026

Citation	:	Imelda., Dewi, I., & Fauzi, KMS.M.A. (2026). Validity And Reliability Analysis of A Contextual Problem-Based Essay Test For Mathematical Critical Thinking On Cubes and Rectangular Prisms . <i>Jurnal PAJAR (Pendidikan dan Pengajaran)</i> , 10(4), 722-733. DOI: http://dx.doi.org/10.33578/pjr.v10i4.1580 .
----------	---	---

INTRODUCTION

In terms of 21st-century skills, students are expected to be proficient in communication, critical thinking, collaboration, and problem-solving (Amina G, 2025; Ardiansyah et al., 2025; Intan et al., 2025; Qiqieh et al., 2026; Qizi, 2024). Critical thinking is a fundamental competency that students must develop (Amelia et al., 2025; Irham et al., 2022; Kharisma et al., 2023; Tandililing et al., 2025). Critical thinking is one of the key competencies students must possess to compete globally, keep up with technological advancements, and handle the complexities of everyday problems. In mathematics education, critical thinking is essential because mathematics emphasizes not only computational skills but also the ability to analyse, evaluate, and draw logical conclusions about problems (Kamila & Mukhlishina, 2026; Lubna et al., 2023; Mutchima et al., 2025; Syaiful et al., 2025; Tambunan et al., 2025; Tanjung et al., 2024). Critical thinking also trains students to think logically and not accept things easily (Deshanty et al., 2023). Critical thinking skills contribute to the development of problem-solving skills (Kuswendi & Maftuh, 2026; Nofiarida, 2023; Rohman et al., 2023; Saadah et al., 2023).

However, the reality is that students' critical thinking skills in mathematics in Indonesia remain relatively low. This is evident from international studies, such as the Program for International Student Assessment (PISA), which show that Indonesian students' ability to solve problems requiring high-level reasoning remains in the lower range (OECD, 2023). One cause of students' low critical thinking skills is that teachers are not yet accustomed to administering critical thinking assessments (Liloi et al., 2025). Teachers are still not accustomed to developing appropriate critical thinking assessment instruments (Hulukati & Rahmi,

2020). Most tests used by teachers still focus on procedural and memorization skills, so they are not yet able to optimally assess students' mathematical critical thinking. Teachers also remain focused on learning outcomes rather than on assessing thinking skills (Admadja & Marpanaji, 2016; Putu et al., 2020).

Furthermore, mathematics instruction often still relies on routine problems that lack real-world context. In fact, the use of contextual problems can help students understand mathematical concepts more meaningfully and encourage them to think critically when solving real-world problems. Therefore, there is a need for a test instrument that assesses mathematical critical thinking skills through contextual problems and measures students' abilities more comprehensively.

Rectangular prisms and cubes are among the geometric concepts most closely related to daily life. Many contextual problems can be developed from this topic, such as calculating the volume of storage boxes, determining the surface area of packaging, and solving problems related to rectangular prisms and cube-shaped buildings. Therefore, the topic of rectangular prisms and cubes has great potential for use in developing a contextual problem-based test instrument to measure students' mathematical critical thinking skills.

Although numerous studies have developed instruments to assess mathematical thinking skills, most have focused on general mathematical achievement, higher-order thinking skills, or problem-solving ability, without specifically integrating contextual problems into the assessment of mathematical critical thinking in geometry. Moreover, previous studies have generally reported validity or reliability separately, while comprehensive evidence encompassing both content validity, construct validity, and reliability for essay-based contextual assessment instruments remains limited. Consequently, there is still a need for a rigorously validated assessment instrument specifically designed to measure students' mathematical critical thinking skills through contextual problems on cubes and rectangular prisms. This study addresses this gap by developing and comprehensively validating an essay-based contextual assessment instrument for geometry learning.

However, the development of contextual problem-based test instruments cannot rely solely on content. A good test instrument must meet validity and reliability criteria. Before a test instrument is used in research or instruction, validity and reliability analyses are required to ensure it accurately and consistently measures students' mathematical critical thinking skills. Validity indicates the extent to which an instrument measures what it is intended to measure, while reliability indicates the consistency of the measurement results. An instrument that is neither valid nor reliable will produce inaccurate data, potentially affecting the quality of research results and the learning evaluation process.

This study was conducted to develop and evaluate an essay-based contextual assessment instrument to measure students' mathematical critical-thinking skills regarding cubes and rectangular prisms. Unlike previous studies that primarily emphasized either validity or reliability of general mathematics assessments, this study provides comprehensive evidence by examining content validity through expert judgment, construct validity through empirical item analysis, and internal consistency reliability within a single instrument development process. The study therefore contributes both theoretically and practically by providing a standardized contextual assessment instrument that aligns mathematical critical thinking with authentic real-life situations. The resulting instrument is expected to support researchers and mathematics teachers in conducting more accurate assessments of students' mathematical critical thinking skills in geometry learning, particularly at the elementary and secondary education levels.

LITERATURE REVIEW

Previous studies have demonstrated growing interest in developing assessment instruments to measure students' critical thinking skills; however, their focus, assessment formats, and psychometric evidence differ considerably. Juliani and Erita (2023) examined the validity and reliability of a multiple-choice mathematics test instrument, whereas Lukman et al. (2023) developed an instrument based on Ennis's FRISCO critical thinking framework. Similarly, Pramarth et al. (2023) emphasized higher-order thinking skills through analysis, evaluation, and creativity, while Dina Putri and Yuanita (2025) developed PISA-oriented assessment items to measure students' critical thinking skills. In addition, Wangsa et al. (2021) focused on critical thinking

assessment in science education, whereas Nada and Yulia (2023) designed essay items based on Facione's critical thinking indicators for the topic of sequences and series. Although these studies make valuable contributions to assessment development, they primarily focus on the critical thinking construct, item validity, or specific assessment formats, without explicitly integrating authentic contextual geometry problems into the assessment design.

Recent international studies have further highlighted the importance of contextual and authentic assessment for evaluating higher-order thinking skills. Authentic assessment enables students to demonstrate reasoning, decision-making, and problem-solving abilities in meaningful real-life situations rather than merely recalling procedural knowledge. Furthermore, contemporary research on educational assessment emphasizes that high-quality assessment instruments should not only demonstrate adequate psychometric properties but also align with authentic learning contexts and the cognitive processes expected in twenty-first-century education. Nevertheless, empirical studies that simultaneously integrate contextual mathematics problems, essay-based assessment, and comprehensive psychometric evaluation remain relatively limited, particularly in geometry learning.

Therefore, a clear research gap exists. Most previous studies have focused on either multiple-choice assessments, PISA-like items, higher-order thinking skills, or general mathematical critical thinking, without specifically developing an essay-based contextual assessment instrument for geometry supported by comprehensive evidence of content validity, empirical item validity, reliability, item difficulty, and discrimination indices. Addressing this gap, the present study develops and evaluates a contextual problem-based essay test designed to measure students' mathematical critical thinking skills on the topics of cubes and rectangular prisms using Facione's critical thinking framework. Accordingly, this study contributes to the growing body of research on mathematics assessment by providing preliminary psychometric evidence for an authentic contextual assessment instrument that may support both classroom assessment practices and future large-scale instrument development.

METHODS

Research Design

This study employed a quantitative descriptive design to evaluate the psychometric quality of a contextual problem-based essay test designed to assess students' mathematical critical thinking skills. The instrument was evaluated using quantitative evidence, including content validity, empirical item validity, internal consistency reliability, item difficulty, and item discrimination indices. This design was considered appropriate because the study focused on describing the measurement characteristics of the developed assessment instrument rather than examining causal relationships among variables.

Research Setting and Participants

The study was conducted at SMPN 31 Medan in March 2026. The participants were 30 ninth-grade students, selected using purposive sampling. The school was chosen because it had implemented the mathematics curriculum covering cubes and rectangular prisms, ensuring that the participants had previously studied the concepts assessed by the instrument. The sample size of 30 students was considered appropriate for the pilot-testing phase of instrument development. In educational measurement, a sample of approximately 30 participants is widely regarded as adequate for preliminary item analysis, enabling researchers to identify problematic items and obtain initial evidence of the instrument's psychometric quality before conducting large-scale validation studies. Therefore, the purpose of this pilot study was to evaluate the instrument's quality rather than to generalize the findings to a broader population. The assessment instrument consisted of five essay items developed based on Facione's critical thinking framework, which includes five indicators: interpretation, analysis, evaluation, inference, and explanation.

Implementation of Action

This study analyzes the validity and reliability of a test instrument for mathematical critical thinking skills, based on contextual problems involving blocks and cubes. The steps taken to achieve the objectives include:

1. Developing a test instrument for mathematical critical thinking skills based on contextual problems by first creating a test instrument blueprint.
2. Preparing a test instrument validation sheet
3. Distributing the test instrument to experts, specifically one lecturer from the Mathematics Education Program at Santo Thomas Catholic University and one doctoral student from the Mathematics Education Program
4. Analyzing the expert validation results and revising and refining the instrument based on the validation findings.
5. After the instrument was declared valid by the experts, the researcher pilot-tested it with students at SMPN 31 Medan.
6. Student answer sheets were collected, scored, and evaluated in accordance with the scoring guidelines.
7. Calculating and analyzing the summary of scores based on students' answers to each question.

Data Collection Techniques

The primary data-collection tool is a test of students' critical-thinking skills in mathematics. The test was developed based on Facione's indicators of mathematical critical thinking, specifically focusing on interpretation, analysis, evaluation, inference, and explanation. The test includes material on cubes and rectangular prisms presented in problem-solving contexts. Below is the test of mathematical critical thinking skills.

Table 1. Critical Thinking Assessment Tool: Cubes and Blocks

Indicator	Question Number	Questions	Cognitive Level
Interpretation	1	A container with a square base with an area of 100 cm^2 , and the height of the container is equal to the length of the sides of the base. Draw the container. Is the container cube-shaped or block-shaped?	C4
Analysis	2	Give the reason.	C4
Evaluation	3	After studying the material on cube volume, Budi gave the following statement: "If I were to make a cube B whose rib length is 2 times the length of Cube A's ribs, then the volume of cube B would be twice the volume of Cube A." Check if Budi's statement is true.	C5
Inference	4	A fish farmer wants to build two fish ponds. The first pool is cube-shaped, and the second pool is block-shaped. The first pond has a side length of 10 m, and the second pond has a length of 20 m and a width of 5 m. If the water from the first pond is drained into the second pond, the second pond will fill. Write down the information on the questions and the calculations so that you can conclude which of the first and second pools requires more materials to make.	C5

Explanation	5	Gita wanted to make 4 block frames of the same size, namely 15 cm x 12 cm x 8 cm. If the price of wire per meter is Rp. 25,000, and Gita has as much money as Rp. 200,000.	C5
		Systematically explain the steps that Gita took to find out whether or not Gita's money was enough to buy the wire.	

The scoring guidelines for the mathematical critical thinking test are organized according to the indicators of mathematical critical thinking: interpretation, analysis, evaluation, inference, and explanation. These are outlined below:

Table 2. Interpretation indicator scoring guidelines

Yes	Indicator	Criteria	Score
1.	Interpretation	Unable to give meaning/interpret data/citation	0
		Interpreting information incorrectly, not according to the meaning/interpretation of the data/situation.	1
		Interpreting information is partly correct; however, identifying relevant data and concepts is correct. There are errors in interpreting the meaning and data/situations.	2
		Interpreting most of the information correctly and identifying relevant data and concepts, as well as meaning/interpreting data/situations, is quite precise, but there are still minor shortcomings in explanation.	3
		Write down all information very precisely, identify relevant data and concepts, and interpret data/situations and meanings correctly and completely.	4
2.	Analysis	Not doing information relationship analysis	0
		Only applying formulas without analysis of the relationships between concepts	1
		Recognizing the relationship between information and information is not precise; the analysis of the relationship between concepts remains weak.	2
		Recognizing the information relationship is mostly correct, the conceptual relationship analysis is quite precise but lacks depth.	3
		Correctly recognize information relationships, compare logically, and provide in-depth analysis with mathematical reasoning	4
3.	Evaluation	Not providing any evidence at all	0
		Providing irrelevant evidence	1
		Presents mathematical proofs, but there are still many errors.	2
		Presents mathematical proofs, but there are still a small number of errors.	3
		Presenting correct mathematical proofs	4
4.	Inference	Unable to draw logical conclusions based on information or calculations	0
		Conclusions are not supported by calculations and explanations of relevant information	1
		Incomplete conclusions or minor errors in reasoning.	2
		The conclusion is logical and supported by calculations, but the explanation of the implications is still lacking.	3
		Draw logical and consistent conclusions, supported by correct calculations, and reflect the demands of the question completely, and explain the implications in detail.	4
5.	Explanation	It does not explain.	0
		Explanations are uncomplicated and difficult to understand	1
		The explanation is less systematic, and there are logical leaps	2

The explanation is quite concise and correct, but the reason is not clear and logical	3
Explain the solution steps in a sequential, systematic, and coherent manner, accompanied by mathematical reasons	4

At the data collection stage, the main data in this study are student answer sheets, scoring guidesheets, and student exam question sheets. Students' answers are corrected by assigning a score to each completed question item.

Content Validity

Before the pilot test, the instrument was evaluated by three experts in mathematics education and educational assessment. The experts assessed each item regarding its relevance to the construct, clarity of wording, appropriateness of the contextual problems, and alignment with the indicators of mathematical critical thinking skills. The degree of agreement among experts was quantified using Aiken's V coefficient, calculated as:

$$V = \frac{\sum s}{n(c - 1)}$$

where s represents the adjusted rating score, n is the number of experts, and c is the number of rating categories. The obtained Aiken's V coefficients ranged from 0.80 to 0.89, indicating high content validity; values of 0.90–1.00 indicate excellent content validity and exceed the recommended acceptance criterion of 0.80. Therefore, all five items were considered appropriate for empirical testing.

Empirical Item Validity and Reliability

Following expert validation, the instrument was administered to the pilot sample to obtain empirical evidence of item quality. Empirical item validity was examined using the Pearson Product-Moment correlation coefficient by correlating each item score with the total test score. Items with correlation coefficients exceeding the critical value at the 5% significance level were considered empirically valid. The instrument's internal consistency reliability was estimated using Cronbach's alpha. According to commonly accepted criteria, a Cronbach's Alpha value of 0.70 or higher indicates satisfactory reliability for educational assessment instruments. In addition to validity and reliability analyses, the psychometric quality of each essay item was further evaluated through item difficulty and item discrimination analyses. The item difficulty index was used to determine the proportion of students who successfully answered each item, whereas the discrimination index measured each item's ability to distinguish between students with high and low mathematical critical thinking skills. These analyses provided comprehensive evidence regarding the quality of the developed assessment instrument.

Data Analysis

Validity was tested using the Pearson product-moment correlation, which was formulated as follows:

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

The correlation coefficient of the calculation results is then interpreted with the following classification:

Table 3. Classification of validity categories

Limitations	Categories
$0.80 < r_{xy} \leq 1.00$	Very High
$0.60 < r_{xy} \leq 0.80$	High
$0.40 < r_{xy} \leq 0.60$	Medium
$0.20 < r_{xy} \leq 0.40$	Low
$0.00 < r_{xy} \leq 0.20$	Very Low
$r_{xy} \leq 0.00$	Invalid

The reliability test is carried out using the Alpha Cronbach formula, as follows:

$$r_{11} = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2} \right)$$

The interpretation criteria regarding correlation coefficients are as follows:

Table 4. Classification of reliability categories

Limitations	Categories
$0.80 < r_{11} \leq 1.00$	Very High
$0.60 < r_{11} \leq 0.80$	High
$0.40 < r_{11} \leq 0.60$	Medium
$0.20 < r_{11} \leq 0.40$	Low
$-1.00 < r_{11} \leq 0.20$	Very Low

RESULTS AND DISCUSSION

The contextual problem-based mathematical critical thinking assessment instrument was first administered to a team of experts in this case, two faculty members to evaluate the content validity of the contextual problems developed. The following is a summary of the evaluation of the contextual problem-based mathematical critical thinking assessment instrument.

Table 5. Aiken's V Coefficient

Aspect	Aiken's V	Interpretation
Test Validity	0,81	High
Content Validity	0,91	Excellent
Construction	0,81	High
Linguistics	0,83	High

The content validity of the developed instrument was evaluated by three experts using Aiken's V coefficient. As presented in Table 5, the obtained Aiken's V values ranged from 0.81 to 0.91, indicating that all assessed aspects met the recommended criterion of 0.80 for acceptable content validity. Specifically, the content validity aspect achieved the highest coefficient ($V = 0.91$), indicating excellent agreement among the experts regarding the relevance of the test items to the construct of mathematical critical thinking skills. The test validity and construction aspects each obtained an Aiken's V coefficient of 0.81, while the linguistic aspect obtained a coefficient of 0.83, all of which fall within the high validity category. These findings indicate that the instrument has appropriate content, well-constructed test items, and clear language, making it suitable for

empirical pilot testing after minor revisions suggested by the experts. Overall, the results demonstrate that the developed contextual problem-based essay test adequately represents the intended construct and provides strong evidence of content validity.

The expert validation results indicate that the developed instrument demonstrates satisfactory content quality across all evaluation aspects, including presentation, content relevance, construction, and language. The highest ratings were obtained for the relevance of the contextual problems to the geometry content and the alignment of the items with the indicators of mathematical critical thinking, indicating strong agreement among the experts that the instrument adequately represents the intended construct. Minor revisions suggested by the experts primarily concerned improving the clarity of instructions, refining several linguistic expressions to avoid ambiguity, and enhancing the consistency of the scoring rubric. These revisions were implemented before the instrument was administered to students, thereby improving its overall quality and usability.

After the expert panel confirmed the test instrument's validity and minor revisions were made, the next step was to pilot it with students. The data used to calculate the validity and reliability of the mathematical critical thinking test instrument were the answer sheets from the pilot test administered to ninth-grade students at SMPN 31 Medan. Before the test was administered to the students, the test instrument had been validated by two lecturers from the Mathematics Education Program at Santo Thomas Catholic University.

After the students answered the test questions, the teacher collected the answers and graded the test according to the established scoring guidelines. The teacher then compiled the students' scores and grades. To determine the instrument's validity with a sample of 30 students, the researcher used the critical r value for a two-tailed test at $\alpha = 0.05$, yielding 0.361. A comparison was then made between the calculated r and the table r . If the calculated r is greater than the table r , the test item is valid; otherwise, it is invalid. The following are the data from the validity calculations for the critical thinking skills test.

Table 8. Test Validity Results

Question number	Correlation (r)	Validity	Categories
1	0,6	valid	Moderate
2	0,695	valid	High
3	0,767	valid	High
4	0,471	valid	Moderate
5	0,457	valid	Moderate

Table 8 shows that there are 2 test items in the high validity category and 3 in the moderate validity category. This means the test instrument is valid and can be used to measure students' mathematical critical thinking skills in solving contextual problems involving cubes and rectangular prisms. The following are the test reliability results, calculated from the validity data for each test item. Reliability testing must be conducted after each test item's validity is calculated. To test reliability, the formula is used. Based on the calculation results, the reliability of the mathematical critical thinking ability test instrument is 0.55, placing it in the moderate category.

Although all five items met the empirical validity criterion, three items demonstrated only moderate validity. This finding suggests that these items measured mathematical critical thinking but were less strongly associated with the overall test score than the other items. One possible explanation is that contextual essay problems generally require students to simultaneously interpret the real-world context, translate it into mathematical representations, and justify their reasoning. Students who had difficulty understanding the context may have produced incomplete responses despite possessing the underlying mathematical knowledge. Consequently, the correlation between these items and the total score became lower than that of the other items.

In contrast, Items 2 and 3 demonstrated high empirical validity because they more effectively distinguished students with higher and lower levels of mathematical critical thinking. These items required

students not only to perform mathematical calculations but also to analyze information, evaluate solution strategies, and provide logical justifications, consistent with Facione's critical thinking framework. Therefore, these items showed stronger relationships with the overall construct being measured. The instrument achieved a Cronbach's Alpha of 0.55, indicating moderate internal consistency. Although this value is below the level generally recommended for large-scale standardized assessments, it is considered acceptable for the preliminary pilot-testing stage of instrument development. At this stage, the primary objective is to identify the instrument's initial psychometric characteristics and to detect items requiring further refinement before wider implementation. Several factors may explain the moderate reliability obtained in this study. First, the instrument consisted of only five essay items. Instruments with a limited number of items generally produce lower internal consistency coefficients because Cronbach's Alpha is influenced by test length. Second, essay questions assess complex cognitive processes, resulting in greater variability in students' responses than objective test items. Third, the relatively small pilot sample (30 students) may have contributed to less stable reliability estimates. Consequently, although the instrument demonstrates adequate reliability for preliminary use, future studies should involve larger samples and additional items to improve its stability and internal consistency.

The findings demonstrate that contextual problem-based essay items can effectively assess mathematical critical thinking by requiring students to integrate conceptual understanding with reasoning in authentic situations. Unlike routine procedural questions, contextual problems encourage students to interpret information, analyze relationships among mathematical concepts, evaluate alternative solution strategies, and justify their conclusions. These cognitive processes closely align with the dimensions of critical thinking proposed by Facione, suggesting that contextual assessment provides a more comprehensive evaluation of students' higher-order thinking skills. The moderate validity observed across several items also highlights the challenges of designing contextual assessment tasks. The quality of contextual scenarios, linguistic clarity, and the cognitive demands imposed by the questions all influence students' responses and, consequently, the psychometric quality of the instrument. Therefore, developing high-quality contextual assessment instruments requires careful consideration of both mathematical content and assessment design principles.

This study contributes to the development of a valid and reliable mathematical critical thinking test instrument based on contextual problems for cube and block material. This study expands upon the research on mathematical critical thinking test instruments conducted by Pramarth et al. (2023), who developed a critical thinking instrument based on Higher Order Thinking Skills (HOT). This study also includes HOTS indicators from Table 1 and is contextual and problem-based. This study also expands upon the research by Juliani and Erita (2023), which only assessed the validity and reliability of a mathematical critical thinking instrument using objective or multiple-choice test items, without specifying the characteristics of the questions. In contrast, the problems designed in this study's instrument are contextual and align with the indicators of critical thinking skills at the Higher Order Thinking Skills (HOTS) cognitive level. Based on this, the resulting instrument can be used by teachers as an alternative evaluation tool to measure the critical thinking skills of junior high school students and to support mathematics learning oriented toward higher-order thinking and mathematical critical thinking through contextual problems.

CONCLUSIONS AND RECOMENDATION

This study demonstrates that the contextual problem-based essay test developed to assess students' mathematical critical thinking on the topics of cubes and rectangular prisms has satisfactory content validity, acceptable empirical item validity, and moderate internal consistency during pilot testing. The instrument also reflects higher-order cognitive processes, particularly at Bloom's taxonomy levels C4 (analysis) and C5 (evaluation), indicating that it can assess both higher-order thinking skills and mathematical critical thinking in authentic, contextual situations. The findings have several practical implications for mathematics education. For teachers, the developed instrument provides a model for designing contextual essay questions that assess not only students' computational abilities but also their reasoning, interpretation, evaluation, and justification of mathematical ideas. Such assessments may support more meaningful classroom evaluation by encouraging

students to connect mathematical concepts with real-life situations. The instrument may also serve as a reference for teachers in developing similar assessments for other mathematics topics aligned with higher-order thinking and critical thinking objectives.

For researchers, this study contributes preliminary psychometric evidence regarding the development of contextual assessments for mathematical critical thinking. The results highlight the importance of integrating expert validation with empirical item analysis during instrument development to ensure that assessment tools accurately measure the intended construct. Nevertheless, this study has several limitations. The empirical evaluation involved a relatively small pilot sample from a single school; therefore, the findings should be interpreted as preliminary evidence rather than definitive validation. Future research should involve larger, more diverse samples from different schools to provide stronger psychometric evidence. Further studies are also recommended to examine construct validity using exploratory or confirmatory factor analysis, improve the instrument's internal consistency by refining or expanding the item set, and investigate its effectiveness in measuring mathematical critical thinking across different mathematical topics and educational contexts.

This study demonstrates that the test instrument for assessing mathematical critical thinking skills based on contextual problems involving cubes and rectangular prisms is valid and reliable. The test instrument also demonstrates higher-order cognitive thinking levels, specifically at levels C4 and C5. This indicates that the test instrument developed can simultaneously assess indicators of higher-order thinking and mathematical critical thinking, with contextual problems as the primary context for the questions. Rather than focusing solely on having students solve higher-order thinking problems, teachers should also learn to develop mathematical critical-thinking test instruments so they can train students with questions designed by the teachers themselves. Teachers can use the instrument in this study as an evaluation tool to assess junior high school students' mathematical critical thinking skills regarding cubes and rectangular prisms. This developed instrument can also support mathematics learning focused on higher-order thinking and critical thinking skills, grounded in contextual problems.

Credit authorship contribution statement

First Author: Methodology, Formal analysis, Data curation, Conceptualization. Second Authors: Resources, Project Administration, Methodology. Last Author: Funding acquisition, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Ethical Declaration

All participants provided informed consent prior to their involvement in the study. They were informed about the study's purpose, procedures, and their right to withdraw at any time without consequence.

References

- Admadja, P. I., & Marpanaji, E. (2016). Pengembangan Multimedia Pembelajaran Praktik Individu Instrumen Pokok Dasar SMK di Bidang Keahlian Kerawitan. 6(2), 173–183. <http://journal.uny.ac.id/index.php/jpv>
- Amelia, R., Waluya, S. B., & Agoestanto, A. (2025). Mathematical Critical Thinking Ability Reviewed from Self-Regulated Learning: Systematic Literature Review. *Anatolian Journal of Education*, 10(1), 91–102. <https://doi.org/10.29333/aje.2025.1017a>
- Amina G, K. (2025). Rethinking Education: The Role of Critical Thinking in Schools. *Newport International Journal of Current Issues in Arts and Management (NIJCIAM)*, 6(1), 112–120. <https://doi.org/10.59298/NIJCIAM/2025/6.1.112120>

- Ardiansyah, A. S., Lestari, E. P., & Ma'ariffatun, K. (2025). Improving Critical Thinking Skills through Mathematics Textbooks Integrated with a Challenge-Based Learning Model in a STEM Context Assisted by the Quizizz Application. *Jurnal Didaktik Matematika*, 12(2), 290–311. <https://doi.org/10.24815/jdm.v12i2.41256>
- Deshanty, N. A., Kartikowati, RR. S., & Syabus, H. (2023). The Use of the Problem Based Learning (PBL) Model in Improving Students' Critical Thinking Skills on Economic Learning Subjects at SMA Negeri 2 Karimun. *JURNAL PAJAR (Pendidikan Dan Pengajaran)*, 7(3), 586–598. <https://doi.org/10.33578/pjr.v7i3.9402>
- Diana, N., Agustinasari, A., Khatimah, H., Ibnuaputra, M., Adiansha, A. A., & Muslim, M. (2025). HOTS Geometry Analysis on Critical Thinking and Numeracy Skills, Reviewed Learning Style and Gender. *Edumatica : Jurnal Pendidikan Matematika*, 15(2), 245–264. <https://doi.org/10.22437/edumatica.v15i2.44262>
- Hulukati, W., & Rahmi, M. (2020). Instrumen Evaluasi Karakter Mahasiswa Program Pendidikan Guru Pendidikan Anak Usia Dini. *Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini*, 4(2), 851. <https://doi.org/10.31004/obsesi.v4i2.468>
- Intan, S., Saputra, E., & Qausar, H. (2025). Pengaruh Pendekatan Contextual Teaching and Learning Berbantuan Video Animasi 3D Berbasis Platagon Terhadap Kemampuan Berpikir Kritis Matematis Siswa. In *Jurnal Pendidikan Matematika Malikussaleh* (Vol. 5, Number 3, pp. 335–363). <https://doi.org/https://doi.org/10.29103/jpmm.v5i3.22888>
- Irham, Tolla, I., & Jabu, B. (2022). Development of the 4C Teaching Model to Improve Students' Mathematical Critical Thinking Skills. *International Journal of Educational Methodology*, 8(3), 493–504. <https://doi.org/10.12973/ijem.8.3.493>
- Juliani, R. P., & Erita, S. (2023). Analisis Validitas dan Reliabilitas Instrumen Penilaian Kemampuan Berpikir Kritis dalam Konteks Sekolah Menengah. *Journal of Educational Integration and Development*, 3(3), 169–179.
- Kamila, K., & Mukhlisina, I. (2026). Qualitative Analysis of the Literacy and Numeracy Habituation Program on the Critical Thinking Skills of Grade III Students at SDN Pendem 01. *Jurnal PAJAR (Pendidikan Dan Pengajaran)*, 10(1), 26–34. <https://doi.org/10.33578/pjr.v10i1.515>
- Kharisma, A. I., Suryaning, A. F., MZ, A., Elisah, A. D. M., Fillah, B. Z., & Widiyanti, W. (2023). The Effectiveness of Hybrid Project-Based Learning Model in Improving Primary Teacher Education Students' Critical Thinking and Creative Thinking Skills. *Primery: Jurnal Pendidikan Guru Sekolah Dasar*, 12(5), 1203–1215. <https://doi.org/10.33578/jpfkip.v12i5.9967>
- Kuswendi, U., & Maftuh, B. (2026). The relationship between critical thinking and problem-solving skills in social studies among primary school students ; *Primery: Jurnal Pendidikan Guru Sekolah Dasar*, 15(2), 354–368. <https://doi.org/10.33578/jpfkip.v15i2.354-368>
- Liloi, O., Rochaminah, S., Hasbi, M., & Baharuddin. (2025). Critical thinking profile of junior high school students in solving function problems, as reflected in their mathematical ability. *Desimal: Jurnal Matematika*, 8(2), 375–390. <https://doi.org/10.24042/djm>
- Lubna, Suhirman, & Prayogi, S. (2023). Evaluation of STEM students' critical thinking in terms of cognitive style through problem-based distance learning. *Journal of Education and E-Learning Research*, 10(3), 557–568. <https://doi.org/10.20448/jeelr.v10i3.4972>
- Lukman, H. S., Setiani, A., & Agustiani, N. (2023). Validitas Instrumen Tes Kemampuan Berpikir Kritis Matematis Berdasarkan Teori FRISCO. *SJME (Supremum Journal of Mathematics Education)*, 7(1), 55–67. <https://doi.org/10.35706/sjme.v7i1.6960>
- Mutchima, P., Leelakitpaisarn, Y., Pijitkamnerd, B., Phiwma, N., & Pantrakool, S. (2025). The effect of metaverse technology on multicultural learning: Strengthening the social attitudes, cultural awareness, and critical thinking skills of secondary school students. *Journal of Education and E-Learning Research*, 12(1), 21–30. <https://doi.org/10.20448/jeelr.v12i1.6329>

- Nada, N. Q., & Yulia, P. (2023). Analisis Instrumen Tes Kemampuan Berpikir Kritis Siswa Pada Pembelajaran Matematika Materi Barisan dan Deret. *Malewa: Journal of Multidisciplinary Educational Research*, 1(02), 18–28. <https://doi.org/10.61683/jome.v1i02.32>
- Nofiarida, N. (2023). Implementation of the Project-Based Learning (PJBL) Model as an Attempt to Improve the Third-grade Students' Critical Thinking Skills at SDN 004 Rantau Kopar. *Primary: Jurnal Pendidikan Guru Sekolah Dasar*, 12(2), 534–541. <https://doi.org/10.33578/jpfkip.v12i2.9674>
- OECD. (2023). *PISA 2022 Results (Volume I)*. OECD. <https://doi.org/10.1787/53f23881-en>
- Pramarth, B. N. I., Ayu Lalita Rathintara, I., Gede Astapa, I., Informatika dan Komputer, F., Teknologi dan Bisnis STIOKOM Bali, I., Raya Puputan No, J., Puri Klod, D., Denpasar Tim, K., & Denpasar, K. (2023). Pengembangan Instrumen Tes Kemampuan Berpikir Kritis Berbasis Higher Order Thinking Skills (HOTS). *Journal on Education*, 06(01), 680–686.
- Putu, N., Krismony, A., Parmiti, D. P., Gusti, I., & Japa, N. (2020). *Pengembangan Instrumen Penilaian untuk Mengukur Motivasi Belajar Siswa SD*. <https://doi.org/10.23887/jippg.v3i2>
- Qiqieh, S., Alkhazalah, M. S., Alwaely, S. A., & Saleem, J. (2026). The implications of generative artificial intelligence on the 21st - century skills of pre-service and in-service teachers: A mixed methods study. *Journal of Education and E-Learning Research*, 13(2), 206–213. <https://doi.org/10.20448/jeelr.v13i2.8807>
- Qizi, S. R. S. (2024). The Role Of Pedagogy In Fostering Critical Thinking Skills in Students. *International Journal of Advanced Scientific Research*, 4(1), 74–77. <https://doi.org/10.37547/ijasr-04-01-12>
- Rohman, F., Azzahra, M., & Supriono, J. (2023). Development of PBL Model-Based Students' Worksheet to Improve Students' Critical Thinking Skill in Phase C. *Primarily: Jurnal Pendidikan Guru Sekolah Dasar*, 12(4), 1084–1098. <https://doi.org/10.33578/jpfkip.v12i4.9862>
- Saadah, A. M., Haryono, H., & Raharjo, T. J. (2023). The Effectiveness of the Discovery Learning Model to Improve Elementary Students' Critical Thinking Skills and Self-efficacy. *Primary: Jurnal Pendidikan Guru Sekolah Dasar*, 12(3), 619–628. <https://doi.org/10.33578/jpfkip.v12i3.9844>
- Syaiful, S., Mukminin, A., & Puspayanti, P. (2025). Using APOS theory in learning basic mathematics to promote students' mathematical critical thinking skills. *Discover Education*, 4(1). <https://doi.org/10.1007/s44217-025-00863-2>
- Tambunan, M., Remigius Abi, A., Sofia Tanjung, D., Sipayung, R., & Janson Silaban, P. (2025). The Influence of the Modified Free Inquiry Learning Model on Students' Critical Thinking Skills in Science Learning in Elementary Schools. *Jurnal PAJAR (Pendidikan Dan Pengajaran)*, 9(2), 144–153. <https://doi.org/10.33578/pjr.v9i2.10068>
- Tandililing, P., Napitupulu, B., Imawan, O. R., & Ismail, R. (2025). Predictive model of mathematical literacy ability, specifically university students' critical thinking skills: A case study on the inappropriateness of multilevel regression. *Journal of Education and E-Learning Research*, 12(3), 507–519. <https://doi.org/10.20448/jeelr.v12i3.7398>
- Tanjung, D., Ambarawati, N., Juliana, & Prayuda, M. (2024). Application of the Problem Based Learning Model to Improve Students. *Learning Outcomes and Critical Thinking Ability on Mathematics Lessons in Primary School*. *Jurnal PAJAR*, 8(3), 307–318. <https://doi.org/10.33578/pjr.v8i3.9660>
- Wangsa, G. N., Dantes, N., & Suastra, I. W. (2021). Pengembangan Instrumen Kemampuan Berpikir Kritis dan Hasil Belajar IPA Kelas V SD Gugus IV Kecamatan Gerokgak. *PENDASI: Jurnal Pendidikan Dasar Indonesia*, 5(1), 139–150.