PAEDAGOGIA
Jurnal Penelitian Pendidikan
http://jurnal.uns.ac.id/paedagogia

DOI : 10.20961/Paedagogia.v28i3.109581
Vol. 28, No. 3, 2025, pp. 412-428
p-ISSN 0126-4109
e-ISSN 2549-6670

Assessing the Quality of Mathematics Test– Enumeration Rules to Measure Students' Computational Thinking Skills

Suparman*, Dadang Juandi, Turmudi, Bambang Avip Priatna Martadiputra, Nana Diana
Department of Mathematics Education, Universitas Pendidikan Indonesia, Bandung, Indonesia

Keywords: *Computational Thinking, Enumeration Rules, Mathematics Test, Reliability, Validity*

Article history
Received: 5 October 2025
Revised: 22 October 2025
Accepted: 22 October 2025
Published: 31 October 2025

*Corresponding Author Email:
arman95@upi.edu

Doi: 10.20961/paedagogia.v28i3.109581

Abstract: Computational thinking (CT), one of the most fundamental thinking processes, empowers students in solving mathematical complex problems. In fact, most Indonesian students, particularly in senior high school, still have poor CT skills in mathematics. The quality of the mathematics test can be one of the potential predictors to describe students' CT skills accurately. The test, however, which involves enumeration rules as a content to measure Indonesian students' CT skills, has not yet been developed. A cross-sectional study using a a descriptive-quantitative method was used to design and produce a qualified mathematics test to measure students' CT skills. Three mathematical-essay problems were assessed by four experts in mathematics education and administered to 35 eleventh-grade students at a public senior high school in Bandung City. A number of quantitative analyses, including Aiken's V, Kendall's W test, Pearson correlation, Cronbach's alpha, discrimination index, and difficulty index, were applied to assess the test's quality. The results of this study demonstrated high content and construct validity, as assessed by experts, as well as strong criterion-related validity based on student performance. High reliability was confirmed through Cronbach's alpha, indicating consistent results across all test items. In addition, item analysis revealed very good discrimination indices and recommended difficulty levels, ensuring the test effectively differentiates student skills. The test provides a model for educators and curriculum designers in mathematics education seeking to improve classroom assessments, particularly in assessing students' CT skills.

How to cite: Suparman, Juandi, D., Turmudi, Martadiputra, B. A. P., & Diana, N. (2025). Assessing the Quality of Mathematics Test– Enumeration Rules to Measure Students' Computational Thinking Skills. *PAEDAGOGIA*, *28*(3), 412-428. doi: 10.20961/paedagogia.v28i3.109581

INTRODUCTION

Computational thinking is a crucial cognitive process that empowers students to analyze, understand, and solve complex problems using systematic and logical reasoning (Suparman, Juandi, Turmudi, et al., 2025). At the senior high school level, this skill is particularly vital for students' success in mathematics, as it underpins a range of mathematical practices, from problem-solving to modeling and algorithmic thinking. When students develop strong computational thinking, they are better equipped to navigate challenging mathematical concepts and apply them in real-world contexts. High mathematics achievement often correlates with students' skills to decompose problems, recognize patterns, generate an automatic formulation, and execute step-by-step reasoning (Suparman et al., 2024; Wing, 2006). In today's rapidly evolving world, where technology and mathematics intersect across industries, fostering computational thinking skills is no longer optional—it's essential. Without it, students may struggle to deal with the demands of modern society (Maharani et al., 2019). Senior high school serves as a formative stage where foundational computational thinking skills are solidified, preparing students for higher education and professional success. Educators, particularly mathematics teachers, must therefore prioritize these skills in their teaching approaches. By focusing on computational thinking, they can build a generation of problem-solvers who excel in mathematics and beyond.

Despite its importance, many senior high school students in Indonesia exhibit poor computational thinking skills in mathematics, particularly in the area of enumeration rules. Numerous studies have highlighted this concern, revealing a consistent trend of low computational thinking skills in this content

(Helsa et al., 2023; Masfingatin & Maharani, 2019; Sukirman et al., 2021; Suparman et al., 2024). For example, research conducted by Maharani et al. (2019) found that students often fail to apply computational thinking skills when solving permutation and combination problems. Similarly, Sukirman et al. (2021) emphasized that students struggle with understanding the fundamental principles underlying enumeration rules, leading to frequent errors. These challenges stem from various factors, including students' limited exposure to problem-solving tasks that foster computational thinking skills. Moreover, ineffective teaching methods that focus on rote learning rather than critical reasoning contribute to the issue (Masfingatin & Maharani, 2019). Students often memorize formulas without truly grasping the underlying concepts, resulting in shallow understanding (Sukirman et al., 2021). Another contributing factor is the lack of integration between enumeration rules and real-world contexts, which makes learning abstract and disconnected. Without authentic engagement and proper scaffolding, students' computational thinking skills remain underdeveloped (Maharani et al., 2019). These issues highlight the urgent need to strengthen students' computational thinking skills, specifically in the context of enumeration rules.

One of the most crucial factors contributing to students' poor computational thinking skills in mathematics is the use of inaccurate instruments to measure these skills. A test that fails to validly and reliably assess computational thinking skills cannot provide an accurate profile of students' skills. Validity ensures that the test measures what it is intended to measure, while reliability guarantees consistent results across different administrations (Nitko & Brookhart, 2014). Moreover, an effective test must consider both the discrimination index—how well it differentiates between high- and low-ability students—and the difficulty index, which indicates how challenging the test items are (Ebell & Friesbie, 1991; Khan et al., 2015). Without these qualities, the results of the mathematics test may misrepresent students' computational thinking skills truly leading mathematics teachers to draw an incorrect profile. For example, a test that is too easy or too difficult fails to provide meaningful insights into students' computational thinking skills. Additionally, poorly designed items can confuse students rather than assess their logical reasoning (Matlock-Hetzel, 1997). Hence, developing a valid and reliable test of computational thinking skills with appropriate discrimination and difficulty indices is essential. Only through rigorous assessment, the teachers can obtain an accurate profile of students' computational thinking skills.

To date, many developmental studies have focused on creating mathematics tests to measure students' problem-solving skills across various mathematics contents (Akveld & Kinnear, 2024; Arriza et al., 2024; Fitrah et al., 2024; Garcia et al., 2025; Kamber Hamzić et al., 2025; Kartianom et al., 2024; Rahmadani & Hidayati, 2023; Shida et al., 2023). These studies have contributed significantly to the field, providing mathematics teachers with valuable tools to assess and enhance student learning. For example, researchers have developed test instruments to assess students' problem-solving in geometry, algebra, and calculus (Garcia et al., 2025; Kartianom et al., 2024; Shida et al., 2023). However, only a few studies have focused on developing tests that specifically measure students' computational thinking skills (Istiqlal et al., 2024; Munawarah et al., 2021). Even among those that exist, the mathematics content embedded in these tests does not include enumeration rules. This condition is a critical gap, as enumeration rules are fundamental to higher-level mathematics, including the addition rule, the multiplication rule, the permutation, and the combination (Jatmiko et al., 2021). The lack of focus on enumeration rules means that mathematics teachers lack appropriate test instruments to assess computational thinking skills in this crucial content. As a result, students' weaknesses remain hidden, and interventions cannot be properly targeted. The present study addresses this gap by developing a test that measures computational thinking skills, specifically in the context of enumeration rules.

The aim of this study is to develop and produce a valid and reliable mathematics test that measures students' computational thinking skills in the context of enumeration rules. By incorporating both discrimination and difficulty indices, the test ensures that it accurately differentiates computational thinking skills among students and provides a meaningful challenge. This test allows mathematics teachers to diagnose students' strengths and weaknesses in computational thinking skills more precisely. Consequently, they can design targeted interventions to support students' learning needs. Moreover, the test can serve as a model for developing similar instruments in other mathematics content, such as algebra, geometry, and statistics. The significant contribution of this study lies in addressing the current gap in assessment tools for computational thinking skills, particularly in enumeration rules. By providing an accurate test instrument, the study empowers mathematics teachers to make a profile of students' computational thinking skills. Overall, this study seeks to make a substantial impact on the quality in assessing students' computational thinking skills in enumeration rules in senior high schools.

Theoretical Framework

Computational Thinking

Computational thinking refers to a problem-solving process, involving a set of cognitive skills rooted in computer science but applicable across disciplines, including mathematics. Wing (2006) popularized the term, defining it as a way of thinking that involves formulating problems in a way that enables computational tools to solve them. Moreover, Wing (2006) explained that computational thinking includes decomposition, abstraction, pattern recognition, and algorithmic thinking. These elements collectively support learners in breaking down complex problems, identifying important details, recognizing patterns, and creating step-by-step solutions. Weintrop et al. (2016) further emphasized that computational thinking is not merely coding but an approach to understanding and solving problems in a structured and logical manner. From the explanations, computational thinking can be presented as an adaptable and strategic method of problem-solving that aligns naturally with mathematical reasoning. In mathematics education, computational thinking enhances students' skills to analyze mathematical problems, structure logical arguments, and execute algorithms (Kaup et al., 2023; Pei et al., 2018). The application of computational thinking in mathematics allows students to construct efficient strategies for solving problems beyond memorization of formulas (Suparman et al., 2024). Hence, integrating computational thinking into mathematics learning enriches students' conceptual understanding and problem-solving capabilities. It positions learners to think critically and reason systematically, which are key to achieving mathematical proficiency.

To measure students' computational thinking skills, a number of indicators have been proposed by educational researchers. Wing (2006) presented four main components, including decomposition, pattern recognition, abstraction, and algorithm. In addition, Zhao and Shute (2019) outline four core components, consisting of decomposition, abstraction, algorithmic thinking, and generalization. Similarly, Angeli et al. (2016) proposed numerous indicators, including problem representation, solution planning, and evaluation of outcomes. Subsequently, Grover and Pea (2013) highlighted the importance of logical reasoning and automation in computational thinking skills. Synthesizing these expert views, computational thinking skills can be measured through students' abilities to break down problems (decomposition), identify patterns (pattern recognition), disregard irrelevant details (abstraction), and construct procedures (algorithmic thinking). Particularly in the present study, a number of indicators, including decomposition, pattern recognition, abstraction, and algorithm, are applied to assess students' computational thinking skills in mathematics. Each indicator represents a distinct yet interrelated dimension of how students apply computational thinking skills to address mathematical problems. By assessing these indicators, teachers can describe how students' computational thinking skills in mathematics. These skills can be operationalized into measurable tasks within mathematics assessments. Therefore, indicators of computational thinking skills serve not only as theoretical constructs but also as practical tools for designing assessments.

Enumeration Rules as One of the Mathematics Contents

Enumeration rules in mathematics refer to principles used to count the number of ways certain events can occur, particularly in problems involving selection and arrangement. This content is a core component of combinatorics and foundational to understanding probability theory. Moreover, the content helps students structure and organize their counting strategies in complex scenarios (Jatmiko et al., 2021). The sub-content discussed in this study includes the addition rule, multiplication rule, permutation, and combination. The addition rule is used when events are mutually exclusive, allowing the total number of outcomes to be found by adding the possibilities. Meanwhile, the multiplication rule applies to compound events, where the total outcomes are found by multiplying the number of ways each event can occur. In addition, Permutation concerns arrangements of objects where order matters, while combination deals with selections where order is irrelevant (Lamanna et al., 2022). Each of these sub-topics involves different reasoning strategies and formulas, requiring precise conceptual understanding. These concepts are often misunderstood due to students' reliance on memorization instead of logical reasoning (Hilda & Siswanto, 2021). Understanding these sub-contents is essential for evaluating students' computational thinking skills in structured counting problems.

There is a strong connection between the contents of enumeration rules and the indicators of computational thinking skills (Matitaputty et al., 2022; Suparman, Juandi, & Turmudi, 2025). Solving problems involving the addition rule or multiplication rule requires students to decompose a scenario into smaller parts and determine the appropriate rule, aligning with the decomposition indicator. Applying permutations or combinations demands abstraction, where students must identify relevant and irrelevant information from a complex problem. To generate accurate solutions, students also need algorithmic thinking, especially when constructing sequences of steps to arrive at answers. For example, in a permutation problem, students must determine how to arrange items systematically, often by applying a formulaic approach. Each of these sub-contents naturally involves identifying patterns and applying logic, both of which are central to computational thinking skills (Lamanna et al., 2022). By integrating enumeration rules into assessments of computational thinking skills, teachers can measure how well students apply logical processes to mathematical contexts. Hence, enumeration rules serve as an ideal domain for operationalizing and assessing students' computational thinking skills in mathematics.

Item Analysis

Developing a high-quality test instrument requires a process known as item analysis, which is used to evaluate the quality of individual test items. This analysis helps ensure that a test is both fair and effective in measuring what it intends to assess. According to Nitko and Brookhart (2014), item analysis involves statistical procedures that provide insight into the validity, reliability, discrimination, and difficulty of test items. Validity determines whether a test measures the intended construct, while reliability concerns the consistency of test results across administrations. In addition, the discrimination index indicates how well an item differentiates between high and low performers, and the difficulty index shows the proportion of students who answer an item correctly (Ebell & Friesbie, 1991). All these components are vital in determining the overall quality and usefulness of a test. Moreover, they help refine test design by identifying items that are too easy, too hard, or not sufficiently discriminating. Hence, well-analyzed items contribute to fair assessments that accurately describe students' competencies, such as students' computational thinking skills in mathematics, particularly in the present study.

Validity and reliability are the two cornerstones of an effective test instrument. Validity refers to the degree to which an instrument measures what it is supposed to measure. According to Nitko and Brookhart (2014), testing validity is a unified concept, including content, construct, and criterion-related validity. In the context of essay tests, content validity ensures that test items align with learning objectives and cover all necessary topics. Additionally, construct validity ensures that test items truly reflect the intended cognitive skills, such as computational thinking. Meanwhile, Nitko and Brookhart (2014) explained that reliability is concerned with the consistency of scores, particularly across different raters or over time. Ensuring reliability involves using clear rubrics that guide scorers in making consistent evaluations. A test with high validity and reliability provides trustworthy results that support meaningful interpretations (Ebell & Friesbie, 1991). Without these, conclusions drawn from the test may be flawed. Thus, rigorous procedures must be in place to assess and maintain validity and reliability, especially in performance-based assessments, such as essay-type tests of computational thinking skills in mathematics. Discrimination and difficulty index are essential in determining the effectiveness of individual test items. The difficulty index measures the challenge of an item, typically calculated as the proportion of students who answer it correctly (Khan et al., 2015). Items that are too easy or too hard may not effectively assess students' skills (Matlock-Hetzel, 1997), particularly in this study, which relates to students' computational thinking skills in mathematics. The discrimination index, meanwhile, reflects an item's ability to differentiate between high-performing and low-performing students (Nitko & Brookhart, 2014). For essay questions, the difficulty index can be inferred from average scores across students, while the discrimination index can be calculated by comparing the scores of top and bottom performers on a specific item. A good item, particularly an essay-type test, should have moderate difficulty and high discrimination, indicating that it challenges students while distinguishing between different skill levels

(Ebell & Friesbie, 1991). These indexes help refine the test to ensure that it serves its purpose effectively. In short, using discrimination and difficulty index enhances the fairness, accuracy, and utility of the assessment, specifically assessing students' computational thinking skills in mathematics.

METHOD

Research Design

This study employed a cross-sectional design using a descriptive-quantitative method to develop a valid and reliable mathematics test for measuring students' computational thinking skills in the topic of enumeration rules. A cross-sectional approach was selected because it enables researchers to collect data at a single point in time, which is ideal for test development and validation purposes (Gall et al., 2014). The descriptive-quantitative method was suitable for analyzing the characteristics of the test, such as validity, reliability, discrimination index, and difficulty index. The method enables a clear and objective description of how the mathematics test performs, based on statistical evidence. It also aligns with common practices in educational test development, particularly when both theoretical and empirical validation are required. The design fits the study's aim, which is not to establish causality but to assess and describe the quality of the test instrument (Creswell & Creswell, 2018). Quantitative data derived from test results and expert validation scores provide measurable indicators of computational thinking skills for the evaluation process. Overall, this design provides a structured and evidence-based approach for developing a mathematics test that accurately assesses computational thinking skills in enumeration rules.

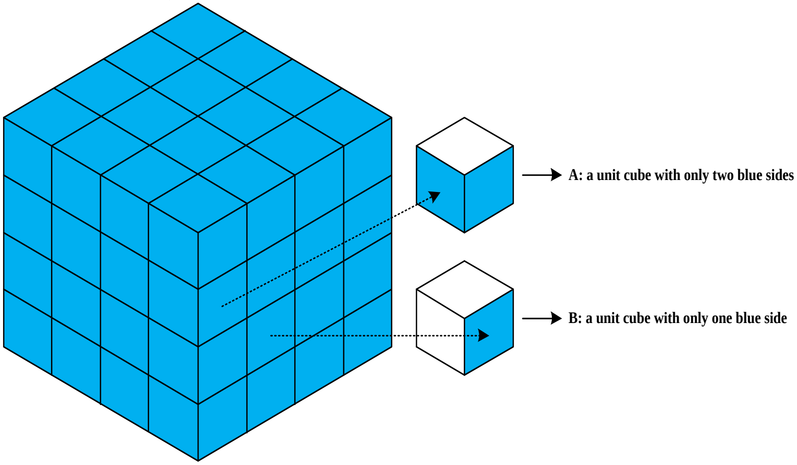
Participant

The respondents of this study were 35 eleventh-grade students enrolled in a public senior high school in Bandung city. These students were selected as the target participants because they had been exposed to the enumeration rules topic in their mathematics curriculum. Their participation provided the necessary data to assess the empirical validity and reliability of the mathematics test. In addition to student respondents, four expert validators were involved to assess the theoretical validity of the test. The validators consisted of two university lecturers and two senior high school teachers, all with a background in mathematics education. Their combined academic and practical expertise ensured a balanced and thorough review of the test instrument. Each validator was asked to evaluate the test based on the content of the enumeration rules, the clarity of the language and answers, and alignment with the indicators of computational thinking skills. Their feedback served as the basis for assessing content and construct validity. By involving both student test-takers and expert validators, the study ensured a robust evaluation process. This dual approach strengthened the credibility of the developed instrument and supported its alignment with both theoretical and practical standards in education.

Instrument and Data Collection

The mathematics test developed in this study consisted of three essay-based problems, including the painted cube problem, the stadium gate problem, and the handshake problem, designed to cover the four sub-topics within enumeration rules, such as addition rule, multiplication rule, permutation, and combination (See Table 1). These sub-topics were selected because they are foundational components in combinatorics and are commonly found in the senior high school mathematics curriculum (Suparman, Juandi, & Turmudi, 2025). The test was explicitly designed to measure four key components of students' computational thinking skills, consisting of decomposition, pattern recognition, abstraction, and algorithm. Each item was constructed to prompt students to apply these skills in solving real-world or semi-contextual mathematical problems. Integrating the indicators of computational thinking skills into the test items was intentional to ensure alignment with the study's objective. The essay format allowed for more in-depth responses and the opportunity to observe students' reasoning processes (Khan et al., 2015). The test was administered to the same 35 eleventh-grade students at a public senior high school in Bandung. Student responses were collected, scored using a rubric based on the indicators of computational thinking skills, and then analyzed statistically.

Table 1. Mathematics Test within the Content of Enumeration Rules

No	Problems	Sub-Content	Computational Thinking Skills	Maximal Score
1	A cube model with dimensions *n* × *n* × *n* is painted blue on all of its outer surfaces. This cube is then cut into unit cubes measuring 1 × 1 × 1. Let *A* be the number of unit cubes that have two blue faces, and *B* be the number of unit cubes that have one blue face (Painted Cube Problem).  A: a unit cube with only two blue sides B: a unit cube with only one blue side	Addition Rule		
	a. Explain your method for determining the values of *A* and *B* for each of the following cube sizes: (1) 3 × 3 × 3 (2) 4 × 4 × 4 (3) 5 × 5 × 5 (4) 6 × 6 × 6 (5) 7 × 7 × 7		Decomposition	10
	b. Based on your answer in part (a), identify and describe a unique pattern for determining the values of *A* and *B*		Pattern Recognition	10
	c. From your findings in part (b), generalize the unique pattern to determine the values of *A* and *B* for a cube model with dimensions *n* × *n* × *n*		Abstraction	10
	d. Explain the steps you would take to determine the value of *B* - *A* if the cube model has a size of 50 × 50 × 50		Algorithm	10
2	A number of people are watching a soccer match at a stadium that has several gates. They enter the stadium through the same gate, but exit through different gates (Stadium Gate Problem).	Multiplication Rule and Permutation		
	a. Explain how you would determine the number of ways the people can enter and exit the stadium if there are: (1) 1 person and 5 gates (2) 2 people and 5 gates		Decomposition	5

No	Problems	Sub-Content	Computational Thinking Skills	Maximal Score
	(3) 3 people and 5 gates (4) 4 people and 5 gates			
	b. Based on your answers in part (a), identify and describe a unique pattern for determining the number of ways people can enter and exit through the stadium gates		Pattern Recognition	10
	c. From your findings in part (b), generalize the pattern to determine the number of ways 3 people can enter and exit the stadium if there are *n* gates		Abstraction	10
	d. Explain the steps you would take to determine the number of ways 3 people can enter and exit the stadium if there are 10 gates		Algorithm	5
3	During an open house event held by Mayor X to celebrate Eid al- Fitr 1446 H, a group of people engage in handshakes. Each person can only shake hands once with each other person, and no one may shake hands with themselves (Handshakes Problem).	Combination		
	a. Explain how you would determine the number of handshakes that occur if there are: (1) 2 people (2) 3 people (3) 4 people (4) 5 people (5) 6 people		Decomposition	5
	b. Based on your answers in part (a), identify and describe a unique pattern for determining the number of handshakes that occur		Pattern Recognition	10
	c. From your findings in part (b), generalize the unique pattern to determine the number of handshakes that occur when there are *n* people at the event		Abstraction	10
	d. Explain the steps you would take to determine the number of handshakes that occur if there are 200 people at the event		Algorithm	5

Research Procedure

To conduct this study, we first designed a mathematics test related to the topic of enumeration rules, including the addition rule, multiplication rule, permutation, and combination. Secondly, four experts in mathematics, including three lecturers and one teacher validated the mathematics test, considering the aspect of content and construct. Thirdly, we revised the mathematics test based on the experts' comments. Fourthly, we administered the mathematics test to 35 eleventh-grade students at a public upper secondary school in Bandung, West Java. Lastly, we collected the data regarding students' computational thinking skills in mathematics and analyzed the data using item analysis.

Data Analysis

To describe the content and construct validity of the mathematics test based on expert assessment, Aiken's V formula was applied. Aiken's validity coefficient helped quantify the degree of agreement among the four validators regarding the relevance and clarity of each item. Aiken (1985) proposed that the Aiken index to calculate the validity size is formulated as:

$$V=\frac{\sum_{i=1}^{n} S_i}{R(c\text{-}1)}$$

Note:

V : Agreement index of experts on item validity

$\sum_{i=1}^{n} S_i$: The sum of scores given by each expert is subtracted from the lowest score within the used category

R : Number of experts

c : Number of categories that can be selected by experts

Moreover, Aiken (1985) categorized the level of validity as: $V < 0.40$ (low validity), $0.40 \le V \le 0.80$ (moderate validity), and $V > 0.80$ (high validity). In addition to Aiken's V, Kendall's W test was conducted to assess the consistency of validators' ratings across all test items. The test provided a statistical measure of inter-rater agreement, ensuring that the validators' assessments were not only valid but also consistent (Rutherford, 2011). For criterion-related validity, Pearson's correlation test was used to examine how well each item's score correlated with overall test performance (Taylor, 1990). This method allowed the researchers to determine whether individual items contributed meaningfully to the overall assessment of computational thinking skills in mathematics. Particularly, Taylor (1990) categorized the coefficient of Pearson's correlation as $r \le 0.35$ (low correlation), $0.36 \le r \le 0.67$ (moderate correlation), $0.67 < r < 0.90$ (high correlation), and $0.90 \le r \le 1.00$ (very high correlation). Furthermore, the Cronbach's Alpha coefficient was employed to analyze the internal consistency reliability of the mathematics test. This measure helped ensure that all items worked together cohesively to assess the same underlying construct (Bland & Altman, 1997). Moreover, Bland and Altman (1997) categorized the level of reliability as $\alpha < 0.50$ (unacceptable reliability), $0.50 \le \alpha < 0.60$ (poor reliability), $0.60 \le \alpha <$ (low reliability), $0.70 \le \alpha < 0.80$ (acceptable reliability), $0.80 \le \alpha \le 0.90$ (high reliability), and $\alpha > 0.90$ (very high reliability). A high Cronbach's Alpha value indicated that the test was reliable and consistent across student responses. These analyses provided a comprehensive evaluation of the validity and reliability of the mathematics test in measuring students' computational thinking skills.

To determine the discrimination level of each item in the mathematics test, the discrimination index proposed by Nitko and Brookhart (2014) was utilized. Particularly, Nitko and Brookhart (2014) proposed that the discrimination index to calculate the discrimination size of each item is formulated as:

$$DI=\frac{2\times(MHG\text{-}MLG)}{MSHG+MSLG}$$

Note:

DI : Difficulty Index

MHG : Mean of High Group

MLG : Mean of Low Group

MSHG : Maximal Score of High Group

MSLG : Maximal Score of Low Group

In a literature, Ebell and Friesbie (1991) categorized the discrimination level as $DI < 0.20$ (poor item), $0.21 \le DI < 0.31$ (marginal item), $0.31 \le DI \le 0.40$ (reasonable good item), and $DI > 0.40$ (very good item). The discrimination index measures how well an item distinguishes between high-performing and low-performing students. For each test item, scores were compared between the top and bottom third of the student group, and the results were used to calculate the index. A higher discrimination index indicates that the item in the mathematics test effectively differentiates between varying levels of student skills in computational thinking. Meanwhile, to describe the difficulty level of each item in the mathematics test, the difficulty index proposed also by Nitko and Brookhart (2014) was applied. Specifically, Nitko and Brookhart (2014) proposed that the difficulty index to calculate the difficulty size of each item is formulated as:

$$\text{DIF}=\frac{\text{MHG}+\text{MLG}}{\text{MSHG}+\text{MSLG}}\times 100$$

Note:
DIF : Difficulty Index
MHG : Mean of High Group
MLG : Mean of Low Group
MSHG : Maximal Score of High Group
MSLG : Maximal Score of Low Group

In a literature, Khan et al. (2015) categorized the difficulty level as *DIF* < 25% (difficult item), *DIF* > 75% (easy item), 25% ≤ *DIF* ≤ 75% (accepted item), and 45% ≤ *DIF* ≤ 60% (recommended item). This index reflects the proportion of students who answered an item correctly, helping to identify whether a question was too easy or too difficult. Both indices are essential for evaluating the quality and fairness of mathematics test items. Items of the mathematics test that are too difficult, too easy, or fail to discriminate effectively were revised or removed. Thus, discrimination and difficulty analyses were integral in refining the mathematics test instrument to measure students' computational thinking skills.

RESULT

The Validity and Reliability of the Mathematics Test

Aiken's V formula was applied to describe the content and construct validity of the mathematics test, as assessed by experts (See Table 2).

Table 2. The Content and Construct Validity of the Mathematics Test

	Assessment Aspect	Expert				Aiken's V	Category
		V1	V2	V3	V4		
1.	The item corresponds to the core subject matter in the question sheet	4	4	4	4	1.00	High Validity
2.	The item used is logical	3	3	3	4	0.75	Moderate Validity
3.	The item contains content that has been studied by students	4	4	4	4	1.00	High Validity
4.	The item is designed to measure students' computational thinking skills	4	4	4	4	1.00	High Validity
5.	Instructions for answering the question are written clearly	4	3	4	4	0.92	High Validity
6.	The question stem is formulated clearly	4	3	4	4	0.92	High Validity
7.	The question stem is free from statements that may cause multiple interpretations	3	3	4	4	0.83	High Validity
8.	The item uses proper and correct Indonesian	4	3	3	4	0.83	High Validity
9.	The item uses communicative sentences	4	3	4	4	0.92	High Validity
10.	The item does not use regionally specific language	4	3	4	4	0.92	High Validity
11.	The language used is appropriate to students' language proficiency level	3	3	4	4	0.83	High Validity
12.	The answer key used is accurate	4	3	4	4	0.92	High Validity
13.	The answer key used is clear	3	3	4	4	0.83	High Validity
14.	The answer key used is logical	4	3	4	4	0.92	High Validity

The results in Table 2 indicate that the mathematics test demonstrates high content and construct validity, as assessed by experts. Of the 14 assessment aspects, 13 aspects achieved high validity, with Aiken's V values ranging from 0.83 to 1.00. Only one item—regarding the logical structure of the question—received moderate validity with an Aiken's V of 0.75. The highest validity scores (1.00) were observed for items that aligned with subject content and measured computational thinking skills. Overall, the test is well-constructed and suitable for assessing students' computational thinking skills in mathematics, particularly in the context of enumeration rules.

In addition to the well-constructed mathematics test, which measures students' computational thinking skills, the results of Kendall's W test indicate that the value of Kendall's W was 0.531, indicating substantial agreement among the experts regarding the mathematics test. Moreover, the results showed that the p-value of Kendall's W test was less than 0.05, indicating significant agreement among the experts regarding the mathematics test. This means the agreement among the validators is not due to chance and is statistically meaningful. Consequently, the validators consistently rated the context and construct aspects of the mathematics test in measuring students' computational thinking skills in the content of enumeration rules.

For criterion-related validity, Pearson's correlation test was used to examine how well each item's score correlated with overall test performance (See Table 3). The test is used to determine whether individual items contributed meaningfully to the overall assessment of computational thinking skills in mathematics.

Table 3. The Results of Pearson's Correlation Test

No	Problem	r Coefficient	Category
1	Painted Cube	0.970**	Very High Validity
	Part (a)	0.956**	Very High Validity
	Part (b)	0.919**	Very High Validity
	Part (c)	0.963**	Very High Validity
	Part (d)	0.955**	Very High Validity
2	Stadium Gate	0.899**	High Validity
	Part (a)	0.614**	Moderate Validity
	Part (b)	0.693**	High Validity
	Part (c)	0.932**	Very High Validity
	Part (d)	0.750**	High Validity
3	Handshakes	0.936**	Very High Validity
	Part (a)	0.927**	Very High Validity
	Part (b)	0.855**	High Validity
	Part (c)	0.844**	High Validity
	Part (d)	0.824**	High Validity

Note: ** < 0.001

The results in Table 3 showed that all items in the mathematics test have statistically significant correlations with total test scores ($p < 0.001$), indicating strong criterion-related validity. The overall problems—*Painted Cube*, *Stadium Gate*, and *Handshakes*—achieved high to very high validity, with correlation coefficients ranging from 0.899 to 0.970. Most individual parts of the problems also demonstrated high or very high validity, except *Stadium Gate Part (a)*, which showed moderate validity ($r = 0.614$). The consistently high *r-values* suggest that each item meaningfully contributes to measuring students' computational thinking skills in mathematics. Therefore, the items of the mathematics test were valid and appropriate for assessing students' computational thinking skills in the topic of enumeration rules.

The Cronbach Alpha coefficient was employed to analyze the internal consistency reliability of the mathematics test (See Table 4). As seen in Table 4, it shows that the mathematics test has strong internal consistency based on Cronbach's alpha values. The *Painted Cube* problem has very high reliability ($\alpha = 0.963$), indicating highly consistent student responses. The *Stadium Gate* problem shows acceptable reliability ($\alpha = 0.707$), while *Handshakes* demonstrates high reliability ($\alpha = 0.861$). The overall reliability of the entire test is very high at 0.908, confirming the instrument's consistency in measuring computational thinking skills in mathematics. These results indicate that the mathematics test is dependable and suitable for measuring students' computational thinking skills in the topic of enumeration rules

Table 4. The results of the Cronbach alpha calculation

No	Problem	α Coefficient	Category
1	Painted Cube	0.963	Very High Reliability
2	Stadium Gate	0.707	Acceptable Reliability
3	Handshakes	0.861	High Reliability
	Overall	0.908	Very High Reliability

The Discrimination and Difficulty Index of Mathematics Test

The results of calculating the discrimination and difficulty index of the mathematics test used for measuring students' computational thinking skills are presented in Table 5.

Table 5. The discrimination and difficulty index of the mathematics test

No	Problem	DI	Category	DIF	Category
1	Painted Cube	0.76	Very Good Item	0.56	Recommended Item
	Part (a)	0.78	Very Good Item	0.60	Recommended Item
	Part (b)	0.68	Very Good Item	0.52	Recommended Item
	Part (c)	0.86	Very Good Item	0.57	Recommended Item
	Part (d)	0.73	Very Good Item	0.53	Recommended Item
2	Stadium Gate	0.58	Very Good Item	0.53	Recommended Item
	Part (a)	0.53	Very Good Item	0.55	Recommended Item
	Part (b)	0.51	Very Good Item	0.60	Recommended Item
	Part (c)	0.63	Very Good Item	0.53	Recommended Item
	Part (d)	0.67	Very Good Item	0.40	Accepted Item
3	Handshakes	0.62	Very Good Item	0.53	Recommended Item
	Part (a)	0.88	Very Good Item	0.54	Recommended Item
	Part (b)	0.64	Very Good Item	0.45	Accepted Item
	Part (c)	0.50	Very Good Item	0.65	Accepted Item
	Part (d)	0.57	Very Good Item	0.47	Recommended Item

Table 5 presents the discrimination and difficulty indices for each item in the mathematics test. All items have discrimination index (DI) values ranging from 0.50 to 0.88, which are classified as "Very Good," indicating that each item effectively distinguishes between high- and low-performing students related to computational thinking skills in mathematics. The highest DI value is found in *Handshakes Part (a)* (0.88), showing strong differentiation ability. The lowest DI value is 0.50 in *Handshakes Part (c)*, which still meets the standard for a good discriminating item. On the other hand, in terms of difficulty index (DIF), most items fall between 0.40 and 0.60, categorizing them as "Recommended" for testing purposes. Only a few items, such as *Stadium Gate Part (d)* and *Handshakes Part (b)* and (c), are labeled "Accepted," suggesting they are slightly more difficult or easier than ideal but still appropriate. The *Painted Cube* problem and its sub-items demonstrate a balanced combination of very good discrimination and recommended difficulty. This reflects that students were sufficiently challenged and that the item performance was consistent across different ability levels. The *Stadium Gate* and *Handshakes* problems also maintained strong metrics, supporting their inclusion in the test. Overall, the mathematics test demonstrates high-quality item construction, effectively assessing students' computational thinking skills in mathematics through valid discrimination and appropriate difficulty.

DISCUSSION

The findings of this study revealed that the mathematics test developed to measure students' computational thinking skills in enumeration rules demonstrates strong validity and reliability. Content and construct validity were confirmed through high Aiken's V scores, with most items achieving coefficients above 0.83, which is categorized as high validity. These results align with the standards of Aiken (1985) for expert judgment-based content and construct validation. Additionally, the consistency among validators, confirmed by Kendall's W = 0.531 and $p < 0.001$, indicated statistically significant agreement, echoing findings by Garcia et al. (2025) on inter-rater reliability in educational assessments. Criterion-related validity was also high, with Pearson's correlation coefficients ranging from 0.614 to 0.970, suggesting that each item contributes meaningfully to the overall test construct. Reliability, assessed using Cronbach's alpha, showed high internal consistency across all items, with an overall coefficient of 0.908. This supports the findings of Tavakol and Dennick (2011), who emphasized the necessity of internal consistency in educational instruments. These results confirm that the mathematics test is both valid and reliable for measuring computational thinking skills in the context of enumeration rules. Together, the findings reinforce the importance of rigorous validation and reliability analysis when designing assessments for computational thinking skills in mathematics.

The discrimination and difficulty indices further support the quality of the test. All items had discrimination indices between 0.50 and 0.88, indicating they effectively differentiate between high- and low-performing students. These values were categorized as "Very Good" according to Nitko and Brookhart (2014), who suggest that a discrimination index above 0.40 reflects a strong item. In terms of difficulty, most items scored between 0.40 and 0.60, placing them in the "Recommended" category, with a few items slightly above or below but still within acceptable bounds. These findings aligned with the work of Ebell and Friesbie (1991), who highlighted the need for balanced difficulty and high discrimination in effective test items. The consistency across items shows that the test not only assesses content knowledge but also the depth of computational thinking skills. Furthermore, essay-based responses allow richer insights into student reasoning, which enhances the interpretation of item performance. These well-constructed items ensure that the test is equitable and functions effectively across a diverse student population. Thus, the performance of the mathematics test in both indices reflects its strength as a tool for diagnostic and evaluative purposes. Overall, the high-quality metrics confirm the effectiveness of the test design.

There is a close relationship between a valid test and a reliable test, particularly in the context of mathematics education. A test with high validity must first ensure that it accurately measures the intended construct—in this case, computational thinking skills. Reliability, on the other hand, ensures consistency across administrations, scorers, or items (Nitko & Brookhart, 2014). According to Matlock-Hetzel (1997), a test cannot be considered valid unless it is also reliable, as inconsistency undermines the accuracy of the measurement. In this study, the high Cronbach's alpha value supports the internal consistency required for validity. Moreover, the strong correlations in criterion validity indicate that reliable measurements align with the test's intended outcomes (Bland & Altman, 1997). Theoretical and statistical validity are therefore closely interlinked, reinforcing each other in practice. As Khan et al. (2015) point out, reliability is a prerequisite for validity, but not all reliable tests are necessarily valid. In this study, the co-occurrence of both validity and reliability confirms that the mathematics test is a sound measurement tool. This reinforces the importance of treating reliability and validity as interdependent in educational assessments.

The relationship between the discrimination index and the difficulty index plays a crucial role in assessing item quality. Items with moderate difficulty often show the highest discrimination, as they are more likely to distinguish between students of varying ability levels (Nitko & Brookhart, 2014). This relationship is consistent with the findings of Kamber Hamzić et al. (2025), who argue that optimal test items fall within a mid-range difficulty level while maintaining high discrimination. In this study, items with difficulty indices between 0.40 and 0.60 achieved the highest discrimination values. For instance, *Handshakes Part (a)* had a high discrimination index of 0.88 and a difficulty index of 0.54, showing strong differentiation. Items that are too easy or too difficult tend to reduce discrimination power, as they cluster responses at the extremes (Kartianom et al., 2024). The observed trends affirm that well-balanced items improve the effectiveness of assessments in identifying student understanding. By ensuring this balance, mathematics teachers can better diagnose specific learning needs and strengths. Hence, maintaining a synergistic relationship between difficulty and discrimination is crucial to designing high-quality tests. The present study successfully achieves this balance across its test items.

A synthesis of the findings showed that a valid and reliable mathematics test with appropriate difficulty and discrimination is essential for measuring computational thinking effectively. As emphasized by Nitko and Brookhart (2014), good assessments must meet technical standards while being fair and educationally meaningful. The test developed in this study meets these criteria, offering both technical quality and pedagogical value. The high validity ensures that the test targets the intended skills, while reliability guarantees consistent measurement. Meanwhile, difficulty and discrimination indices confirm that the test challenges students appropriately and distinguishes varying levels of mastery. These combined attributes enhance the function of the mathematics test as both an evaluative and formative tool (Ebell & Friesbie, 1991). According to Khan et al. (2015), such balanced tests contribute to a more equitable education system by accurately identifying students' learning profiles. The inclusion of essay-based items further strengthens the ability of the mathematics test to measure students' computational thinking skills. Therefore, the test serves as a comprehensive instrument for gauging students' computational thinking skills in the context of enumeration rules.

Implications for the Field of Mathematics Education

Theoretically, this study contributes to a deeper understanding of how computational thinking skills can be assessed through well-designed mathematics tests. It supports the framework proposed by Zhao and Shute (2019), which emphasized decomposition, abstraction, pattern recognition, and algorithmic thinking as core elements of computational thinking skills. The findings confirm that these indicators can be effectively measured using a structured essay-based test in mathematics. Additionally, the study affirms that theoretical constructs in computational thinking can be translated into measurable performance indicators. This bridges the gap between theory and classroom application, particularly in secondary-level mathematics. The study also reinforces the importance of using psychometric standards in test development. By applying classical test theory, such as item analysis and validity-reliability metrics, the research validates theoretical models in a practical educational context. These insights extend the academic discussion around the operationalization of computational thinking skills in school assessments. Thus, the study provides theoretical groundwork for future research aimed at aligning assessment design with the development of 21st-century skills. It deepens the discourse around cognitive assessment in mathematics education.

Practically, this study offers valuable insights for teachers, test developers, and curriculum planners in mathematics education. The test provides a concrete example of how to assess computational thinking skills through contextual and conceptually rich problems. It also serves as a model for integrating content and construct, and criterion-related validity, reliability, difficulty, and discrimination into assessment design. Mathematics teachers can adapt similar approaches to evaluate students' computational thinking processes rather than just final answers. Furthermore, this approach supports differentiated instruction by identifying students' specific strengths and weaknesses. Educational institutions can use such instruments for diagnostic, formative, or even summative purposes. The test format encourages computational thinking and deeper engagement with mathematical content. By aligning assessments with computational thinking, mathematics teachers promote higher-order thinking skills essential for academic and real-life problem-solving. The study thus contributes directly to improving classroom assessment practices regarding students' computational thinking skills in the topic of enumeration rules.

Limitations and Suggestions

This study, unfortunately, has several limitations that should be acknowledged. Firstly, the sample size was limited to 35 eleventh-grade students from a single public senior high school in Bandung, which may limit the generalizability of the findings. A larger and more diverse sample could strengthen the robustness of the results. Secondly, the validators involved were limited to two lecturers and two teachers, potentially narrowing the range of perspectives. Thirdly, only essay-type questions were used, which may not fully capture the range of computational thinking skills across all learning styles. Fourthly, while the test targeted four computational thinking indicators, other relevant dimensions, such as debugging or optimization, were not included. Fifthly, this study employed classical test theory only; incorporating modern psychometric methods, such as item response theory, could provide deeper insights. Additionally, the cultural and linguistic context of the mathematics test may limit its applicability outside the

Indonesian educational system. These limitations suggest areas for caution in applying the test broadly. Recognizing these constraints is important for interpreting the study's results accurately. Future studies should address these issues to enhance the reliability and utility of mathematics assessments.

To address these limitations, future research should consider expanding the sample to include students from different schools, regions, and educational backgrounds. This would help determine whether the test performs consistently across various populations. Involving a broader panel of validators, including international experts, could also enrich the content validation process. Additionally, future tests could include multiple item formats, such as multiple-choice, short answer, or project-based tasks, to capture diverse cognitive processes. Researchers should also explore more advanced psychometric models like item response theory to refine item analysis. Furthermore, integrating other components of computational thinking—such as evaluation, debugging, and automation—would create a more comprehensive assessment tool. Longitudinal studies could examine the test's impact on students' learning progress over time. Adapting the instrument for use in digital or online formats could also increase accessibility. Finally, cross-cultural validation would allow the test to be adapted for broader international use. These suggestions aim to enhance the depth, applicability, and impact of future mathematics assessment research.

CONCLUSION

In conclusion, this study successfully developed and produced a mathematics test designed to measure students' computational thinking skills in the topic of enumeration rules. The test showed high content and construct validity through expert assessment, as well as strong criterion-related validity based on student performance. High reliability was confirmed through Cronbach's alpha, indicating consistent results across all test items. Item analysis revealed very good discrimination indices and recommended difficulty levels, ensuring the test effectively differentiates student abilities. The balance of difficulty and discrimination affirms the test's capacity to assess computational thinking with fairness and precision. Relationships between validity and reliability were confirmed, emphasizing the importance of both in educational measurement. Theoretical frameworks on computational thinking were successfully operationalized through essay-based tasks. The findings contribute to the literature on how to assess computational thinking skills in mathematics. Practically, the test provides a model for educators and curriculum designers in mathematics education seeking to improve classroom assessments. While the study had limitations in scope and methodology, it provides a strong foundation for further research. Future studies should broaden the sample and incorporate diverse item formats and psychometric techniques. Enhancing content coverage and cultural adaptability would further strengthen the tool's utility. The integration of classical test theory with real-world teaching needs demonstrates how theory and practice can be aligned. Ultimately, this study underscores the importance of designing assessments that are both technically sound and educationally meaningful. It contributes to more effective and equitable mathematics education by equipping teachers with tools to evaluate and foster computational thinking skills.

ACKNOWLEDGMENT

The authors are greatly grateful to the educational authority at a public senior high school in Bandung city, which permitted us to administer the mathematics test and collect the data regarding students' computational thinking skills in the topic of enumeration rules.

FUNDING INFORMATION

We are also thankful to the Indonesian Endowment Fund of Education (LPDP) for partial financial support (Granted Number: SKPB-3939/LPDP/LPDP.3/2023).

ETHICAL APPROVAL STATEMENT

This developmental study has been approved by the Faculty of Mathematics and Science Education, Universitas Pendidikan Indonesia, to be conducted at a public senior high school in Bandung, West Java, Indonesia (Number: B-3517/UN40.A4/PK.03.03/2025).

AUTHORS' CONTRIBUTION

All authors contributed to this study. Suparman: designing the mathematics test instrument, and writing the original manuscript; Nana Diana: administering the mathematics test and collecting the data; Bambang Avip Priatna Martadiputra: analyzing the data and interpreting the findings; Dadang Juandi and Turmudi: reviewing and editing the manuscript.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings, educational and psychological measurument. *Educational and Psychological Measurement*, *45*(1), 131–142.

Akveld, M., & Kinnear, G. (2024). Improving mathematics diagnostic tests using item analysis. *International Journal of Mathematical Education in Science and Technology*, *55*(10), 2478–2505. https://doi.org/10.1080/0020739X.2023.2167132

Arriza, L., Retnawati, H., & Ayuni, R. T. (2024). Item analysis of high school specialization mathematics exam questions with item response theory approach. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *18*(1), 151–162. https://doi.org/10.30598/barekengvol18iss1pp0151-0162

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, *314*, 571–572.

Creswell, W. J., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative adn Mixed Methods Approaches*. Sage Publications Inc.

Ebell, R. L., & Friesbie, D. A. (1991). *Essentials of educational measurement*. Prentice-Hall, Inc. https://ebookppsunp.files.wordpress.com/2016/06/robert_l-ebel_david_a-_frisbie_essentials_of_edbookfi-org.pdf

Fitrah, M., Sofroniou, A., Ofianto, Judijanto, L., & Widihastuti. (2024). Reliability and separation index analysis of mathematics questions integrated with the cultural architecture framework using the Rasch model. *Journal of Education and E-Learning Research*, *11*(3), 499–509. https://doi.org/10.20448/jeelr.v11i3.5861

Gall, J. P., Gall, M. D., & Borg, W. R. (2014). *Applying educational research: How to read, do, and use research to solve problems of practice*. Longman publishing Inc.

Garcia, M. L. B., Santos, K. C. P., & Vistro-Yu, C. P. (2025). Comparing two psychometric approaches: The case of item analysis for a classroom test in mathematics. *International Journal of Education and Practice*, *13*(1), 327–344. https://doi.org/10.18488/61.v13i1.4060

Grover, S., & Pea, R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Researcher*, *42*(1), 38–43. https://doi.org/10.3102/0013189X12463051

Helsa, Y., Suparman, Juandi, D., Turmudi, & Ghazali, M. B. (2023). A meta-analysis of the utilization of computer technology in enhancing computational thinking skills: Direction for mathematics learning. *International Journal of Instruction*, *16*(2), 735–758. https://doi.org/10.29333/iji.2023.16239a

Hilda, A. M., & Siswanto, R. D. (2021). Android application development: Permutation of the same elements based on realistic mathematics education. *Mathematics Teaching-Research Journal*, *13*(4), 170–180.

Istiqlal, M., Istiyono, E., Widihastuti, Sari, D. K., Danni, R., & Safitri, I. (2024). Construction of mathematics cognitive test instrument of computational thinking model for Madrasah Aliyah students. *Nazhruna: Jurnal Pendidikan Islam*, *7*(2), 475–492. https://doi.org/10.31538/nzh.v7i2.4425

Jatmiko, M. A., Herman, T., & Dahlan, J. A. (2021). Desain didaktis materi kaidah pencacahan untuk Siswa SMA kelas XI [Didactical design of enumeration rules for eleventh-grade students]. *Hipotenusa Journal of Research Mathematics Education (HJRME)*, *4*(1), 35–54. https://doi.org/10.36269/hjrme.v4i1.464

Kamber Hamzić, D., Trumić, M., & Hadžalić, I. (2025). Construction and analysis of test in triangle and circle trigonometry. *International Electronic Journal of Mathematics Education* , *20*(1), 1–9. https://doi.org/10.29333/iejme/15734

Kartianom, K., Retnawati, H., & Hidayati, K. (2024). Assessing the fairness of mathematical literacy test in Indonesia: Evidence from gender-based differential item function analysis. *Journal of Pedagogical*

Research, *8*(3), 191–208. https://doi.org/10.33902/JPR.202426420

Kaup, C. F., Pedersen, P. L., & Tvedebrink, T. (2023). Integrating computational thinking to enhance students' mathematical understanding. *Journal of Pedagogical Research*, *7*(2), 127–142. https://doi.org/10.33902/JPR.202318531

Khan, G.-A. N., Ishrat, N., & Khan, A. Q. (2015). Using Item analysis on essay types questions given in summative examination of medical college students: Facility value, Discrimination index. *International Journal of Research in Medical Sciences*, *3*(1), 178–182. https://doi.org/10.5455/2320-6012.ijrms20150131

Lamanna, L., Gea, M. M., & Batanero, C. (2022). Do secondary school students' strategies in solving permutation and combination problems change with instruction? *Canadian Journal of Science, Mathematics and Technology Education*, *22*(3), 602–616. https://doi.org/10.1007/s42330-022-00228-z

Maharani, S., Kholid, M. N., Pradana, L. N., & Nusantara, T. (2019). Problem-solving in the context of computational thinking. *Infinity: Journal of Mathematics Education*, *8*(2), 109–116.

Masfingatin, T., & Maharani, S. (2019). Computational thinking: Students on proving geometry theorem. *International Journal of Scientific and Technology Research*, *8*(9), 2216–2223.

Matitaputty, C., Nusantara, T., Hidayanto, E., & Sukoriyanto. (2022). Examining the pedagogical content knowledge of in-service mathematics teachers on the permutations and combinations in the context of student mistakes. *Journal on Mathematics Education*, *13*(3), 393–414. https://doi.org/10.22342/jme.v13i3.pp393-414

Matlock-Hetzel, S. (1997). Basic Cconcepts in item and test analysis. *Educational Research Association*, *1*, 1–15.

Munawarah, Thalhah, S. Z., Angriani, A. D., Nur, F., & Kusumayanti, A. (2021). Development of instrument test computational thinking skills IJHS/JHS based RME approach. *Mathematics Teaching-Research Journal*, *13*(4), 202–220.

Nitko, A. J., & Brookhart, S. M. (2014). Educational Assessment of Students Sixth Edition. In *Pearson New International Edition*. Pearson Education, Inc.

Pei, C. (Yu), Weintrop, D., & Wilensky, U. (2018). Cultivating computational thinking practices and mathematical habits of mind in lattice land. *Mathematical Thinking and Learning*, *20*(1), 75–89. https://doi.org/10.1080/10986065.2018.1403543

Rahmadani, N., & Hidayati, K. (2023). Quality of mathematics even semester final assessment test in class VIII using R program. *Jurnal Pendidikan Matematika*, *17*(3), 397–416.

Rutherford, A. (2011). *ANOVA and ANCOVA: A GLM approach*. John Willey & Sons, Inc.

Shida, N., Abdullah, A. H., Osman, S., & ... (2023). Validation of mathematics test to assess polytechnic students' problem solving. *JPM: Jurnal Pendidikan Matematika*, *17*(2), 265–278. https://ejournal.unsri.ac.id/index.php/jpm/article/view/20171

Sukirman, S., Ibharim, L. F. M., Said, C. S., & Murtiyasa, B. (2021). A strategy of learning computational thinking through game based in virtual reality: Systematic review and conceptual framework. *Informatics in Education*, *21*(1), 179–200. https://doi.org/10.15388/infedu.2022.07

Suparman, Juandi, D., & Turmudi. (2025). Development of Ucing Sumput as a digital educational game to enhance students' mathematics achievement. *Mathematics Teaching-Research Journal*, *17*(2), 7–35.

Suparman, Juandi, D., Turmudi, & Wahyudin. (2024). Computational thinking in mathematics instruction integrated STEAM education : Global trend and students ' achievement in the last two decades. *Beta: Jurnal Tadris Matematika*, *17*(2), 101–134. https://doi.org/10.20414/betajtm.v17i2.643

Suparman, S., Juandi, D., Turmudi, T., Martadiputra, B. A. P., Helsa, Y., Masniladevi, M., & Suherman, D. S. (2025). Computational thinking in mathematics instruction integrated to STEAM education: A systematic review and meta-analysis. *TEM Journal*, *14*(1), 949–963. https://doi.org/10.18421/TEM141-84

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical*

Sonography, *6*(1), 35–39. https://doi.org/10.1177/875647939000600106

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, *25*(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5

Wing, J. M. (2006). Computational thinking. *ACM SIGCSE Bulletin*, *39*(1), 195–196. https://doi.org/10.1145/1227504.1227378

Zhao, W., & Shute, V. J. (2019). Can playing a video game foster computational thinking skills? *Computers and Education*, *141*, 1–40. https://doi.org/10.1016/j.compedu.2019.103633