




## Applying the Rasch model to measure students' critical thinking skills on the science topic of the human circulatory system

Sigit Sujatmika<sup>1</sup> , Sutarno<sup>1\*</sup> , Mohammad Masykuri<sup>1</sup> , Baskoro Adi Prayitno<sup>1</sup> 

<sup>1</sup> Sebelas Maret University, Surakarta, INDONESIA

Received 14 January 2025 ▪ Accepted 05 March 2025

### Abstract

Critical thinking (CT) is essential in science education to enable students to deeply understand scientific concepts and apply their knowledge to solve complex real-world problems. Despite its importance, a notable gap remains in the availability of instruments designed to measure CT in specific science topics. This study addresses this gap by testing the accuracy of an instrument for assessing secondary students' CT related to the human circulatory system, focusing on six APA-defined indicators: interpretation, analysis, inference, evaluation, explanation, and self-regulation. The study was conducted in Bantul Regency, Yogyakarta, Indonesia, and involved 445 8th-grade students from urban, middle, and rural schools. Data were collected through an online test administered in collaboration with teachers after students had completed the topic. Using the Rasch model for data analysis improved the accuracy and consistency of the instrument. Results showed that self-regulation was CT subskill with the lowest mean score, highlighting it as a priority for development. At the same time, interpretation had the highest mean score, particularly in the level 4 rating category, making it the most mastered skill. These findings underscore the need for educators to develop targeted learning strategies that enhance CT skills and adapt them to other science topics with similar complexity to the human circulatory system.

**Keywords:** Rasch model, critical thinking skills, human circulatory system

## INTRODUCTION

Critical thinking (CT) skills are essential for navigating information, evaluating data accuracy, and making decisions based on logic and ethics. As the global community faces challenges such as disinformation and complex social and environmental issues, CT empowers individuals to become resilient problem solvers and responsible global citizens, mindful of the impact of their decisions (Lombard et al., 2020; Marin & Halpern, 2011; Sujatmika et al., 2022; Trilling & Fadel, 2009; Zenker, 2018). CT has been recognized as one of the essential skills for the 21<sup>st</sup> century and is considered a core skill for 2030 by the OECD (Bao & Koenig, 2019; Liu & Pásztor, 2022). The importance of CT is particularly evident in the ability to think analytically, logically, and reflectively when evaluating different arguments and available information (Dwyer et al., 2014; Howard et al., 2015). In education, especially in science learning, CT has gained

increasing prominence, as students are expected to understand scientific concepts and apply their knowledge to solve complex problems (García-Carmona, 2023; Ma et al., 2023). CT in science equips students with the skills to engage in essential scientific processes, such as identifying problems, formulating hypotheses, and testing solutions (Kuhn, 2019; OECD, 2019). CT is essential not only in the classroom for academic success but also in everyday life, particularly in dialogical interactions, decision-making, and problem-solving (Franco et al., 2018).

Furthermore, CT enables students to develop a deeper understanding of scientific concepts and apply their knowledge to real-world contexts (Forawi, 2016; Liu & Pásztor, 2022). Consequently, CT has become a central focus of educational research, with numerous studies exploring strategies to develop and assess these skills at various educational levels (Castro, 2009; Davies, 2015; van Laar et al., 2020). This article examines CT in

### Contribution to the literature

- Addressing the gap in literature by developing and testing a valid and reliable instrument to measure students' CT skills on a specific topic, i.e., the human circulatory system.
- Contributes to measurement methodology in science education, providing a better foundation for future research in assessing CT skills in science topic areas.
- Provides practical direction for educators to design more focused learning strategies, enhance students' specific weaknesses, and strengthen the mastery of CT skills.

the context of the human circulatory system among secondary school students, as its complexity challenges students to analyze, evaluate, and synthesize information. As highlighted by Cardaba (2024), scientific literacy is closely linked to CT, serving as a civic competence that enables individuals to think rationally about socioeconomic or personal issues. A scientifically literate person can differentiate ideas, analyze data, and apply scientific knowledge to address complex situations, underscoring the importance of fostering CT through science education.

CT has been defined by various scholars, each emphasizing complementary aspects of the concept. Facione (1990) adds that CT involves interpretation, analysis, evaluation, and inference based on relevant evidence. Ennis (1991) describes CT as reflective and logical thinking that verifies beliefs and actions. Paul (1992) views it as a mental discipline that supports independent thinking and rational decision-making, whereas Halpern (1998) underscores its importance in making better decisions, especially in complex situations. Bailin (2002) defines CT as thinking that meets standards of adequacy and accuracy, while Lai (2011) highlights the ability to recognize problems and make decisions based on valid information. Paul and Elder (2014) stress that CT involves reasoned judgment and considering diverse perspectives. Brookfield (2017) adds that it requires questioning foundational assumptions and considering the social context that influences our views. These definitions highlight that CT involves evaluating evidence, considering different perspectives, and making more effective and rational decisions. Several definitions of CT converge in the consensus outlined in the APA Report (Facione, 2020), which identified key components of CT. According to this consensus, CT encompasses six core skills:

- (1) interpretation,
- (2) analysis,
- (3) inference,
- (4) evaluation,
- (5) explanation, and
- (6) self-regulation.

We used the APA Report as the foundation for determining CT because the project involved 46 experts from philosophy, psychology, education, social sciences, and other disciplines, all selected for their specialized

knowledge in CT and education (Dwyer et al., 2017). The APA Report also provides a clear framework for CT, outlining its components and the corresponding indicators for each skill.

Although substantial research has been conducted on the importance of CT, a clear gap remains in the literature regarding the development of tools to measure this skill, particularly in specific science topics such as "the human circulatory system." Santos (2017) has reviewed the role of CT in science education but has not addressed the measurement instruments for assessing this skill. At the university level, Tiruneh et al. (2017) developed and validated a CT test on physics in electricity and magnetism. Mapeala and Siew (2015) created a CT test for science at the elementary school level. According to the systematic literature review by Hakim dan Talib (2018), while studies on CT in science exist, the assessment tools used are general instruments and focused on broad contexts. Abosalem (2015) highlights that this gap signals the need to develop more specific instruments to measure CT within secondary science education. A more recent study by Schwichow et al. (2016) further emphasizes the need to develop accurate measures in science education, reinforcing the demand for more focused, content-based instruments. Therefore, this study seeks to fill this gap by developing more accurate and context-specific instruments, particularly for topics such as the human circulatory system.

In this study, we tested the accuracy of an instrument designed to measure secondary school students' CT skills on the topic of the human circulatory system. The definition and indicators of CT skills are based on the APA Report, which includes interpretation, analysis, inference, evaluation, explanation, and self-regulation. This research aims to fill the existing literature gap by providing valid and reliable instruments to assess CT skills in the context of specific science topics while offering teachers guidance in developing CT-based learning strategies in the classroom.

### Research Aim and Research Questions

We aimed to validate and assess the reliability of CT instrument on "the human circulatory system" through a field trial, using the Rasch model analysis to provide insights into the instrument's effectiveness. Specifically, we formulated the following research questions:

**Table 1.** Demography of the participants

No	Area	Gender		Total number of students	Percentage (%)
		Male	Female		
1	Urban	70	103	173	38.88
2	Mid-range	74	73	147	33.03
3	Rural	59	66	125	28.09
Total		203	242	445	100

1. How is the description of students' CT skills in the aspects of interpretation, analysis, evaluation, inference, explanation, and self-regulation on the material of the human circulatory system?
2. How is the level of difficulty of items in each aspect of students' CT measured using the Rasch model?
3. How does the polytomous scale category work to measure students' CT skills in each aspect?
4. How do students' CT skills manifest in the aspects of interpretation, analysis, evaluation, inference, explanation, and self-regulation within the topic of the human circulatory system?

## RESEARCH METHODOLOGY

### Research Design

We used a quantitative method with a survey model to measure CT skills of junior high school students on science topics. The quantitative method was chosen because it allows accurate numerical analysis to describe the level of CT skills in different groups of students. The stratified random sampling technique was determined based on the location of schools in the Bantul Regency so that the sample could be representative of the student population. Bantul is located in Yogyakarta Province, while Yogyakarta is known as one of the education centers in Indonesia. The survey was conducted from September to October 2024.

Experts validated CT instruments to ensure theoretical rigor and alignment with the intended learning objectives. Following field testing, researchers employed the Rasch model for data analysis to enhance both the instrument's accuracy and the consistency of measurement results. This model also enables item calibration, allowing for a more valid interpretation of research findings.

### Participants

This study involved public junior high schools in Bantul Regency, Yogyakarta, Indonesia, categorized into urban, mid-range, and rural areas based on population density, economic activity, and infrastructure, following United Nations (2018) urbanization criteria. Urban areas (3 districts) had high density and developed infrastructure; mid-range areas (5 districts) had moderate density and mixed economies; and rural areas (9 districts) had low density and agriculture-based

economies. Data were sourced from Indonesia's Central Bureau of Statistics. Of the 47 public schools under the Indonesian Ministry of Education, one school from each area category was selected using stratified random sampling to ensure proportional representation and enhance generalizability. This excludes private schools and public schools under other ministries. The categorization aimed to examine disparities in facilities, resources, and learning support across regions, even with universal Internet coverage. The heterogeneous respondent data enabled a realistic analysis, with rural schools expected to have more limited learning support compared to urban and mid-range areas.

The participants consisted of 8th-grade students, both male and female, with a slightly higher number of female students. All participants had completed the study of the human circulatory system in their respective schools by the time of the test, ensuring that their understanding of the concepts would influence the quality of their answers and reduce the risk of careless responses. To support data validity, the researcher collaborated with teachers and integrated the test as a form of practice. The demography of the participants is shown in **Table 1**, with the urban group having the highest number of participants, though the differences among groups were not substantial. Similarly, there was no significant difference in participant numbers based on gender.

The availability of heterogeneous respondent data allows for more realistic analysis results. The total number of students involved was 445. Although it was considered practice, students were not forced to do the test; they could choose to accept it or withdraw from it. Data confidentiality and anonymity were maintained, and respondents knew their data would be used for research.

### Data Collection

Data on students' CT were collected using an online test via Google Form, chosen for its time efficiency and feasibility without compromising the quality of the measurement tool. Students were already familiar with online applications, making this method suitable for classroom activities. Science teachers at each school assisted in the data collection process, with test administration timed to coincide with the completion of the circulatory system topic in each class. Teachers facilitated test distribution, addressed questions during testing, and supervised students as they completed the

**Table 2.** The relationship between CT and the topic of the human circulatory system

Core skill	Sub-skill	Sub-topics of the human circulatory system		
		Components of the circulatory system	Circulatory mechanism and its regulation	Disorders, diseases, and keeping the circulatory system healthy
Interpretation	Categorize	1A & 1B		14A
	Clarify meaning	2A & 2B	10A	13A & 15B
Analysis	Examine ideas	4A & 4B	11A	
	Identify arguments	3A & 3B	12A	
Inference	Query evidence		5A & 5B	
	Justified conclusions	9A & 6B	14B	15A, 10B, & 12B
Evaluation	Assess the quality of arguments	6A & 7B	16A, 19A, & 16B	11B, 13B, & 18B
Explanation	Justify procedures		7A & 8B	17A & 17B
	Present arguments	8A & 9B	19B	18A & 9B
Self-regulation	Self-correction		20A & 20B	

**Table 3.** Example of CT test on explanation

Item indicator	Form of the question
Provide students with an explanation of changes in blood pressure due to physical activity that is supported by data. Students evaluate the quality of the explanation and determine the correct answer.	<p>Consider the following blood pressure data measured on a student under two different conditions.                      Resting condition: Systolic 110 mmHg &amp; diastolic 70 mmHg                      Post-exercise condition: Systolic 140 mmHg &amp; diastolic 90 mmHg</p> <p>The correct explanation of the blood pressure data is ...</p> <p>A. Blood pressure decreases after exercise.                      B. Blood pressure does not change after exercise.                      C. Exercise does not affect blood pressure.                      D. Blood pressure increases after exercise.</p> <p>Reason:                      A. Because exercise increases the oxygen demand, blood pressure rises.                      B. Because exercise decreases oxygen demand, blood pressure decreases.                      C. Because the body has a mechanism for keeping blood pressure constant.                      D. Because physical activity does not affect the cardiovascular system.</p>

test independently. Each student took the online test only once to maintain question confidentiality. Before starting, teachers provided instructions to ensure all students understood the task, and the test was administered within a 60-minute time frame to maintain consistency and focus.

**CT Instrument**

CT measurement technique employed multiple-choice and essay questions, totaling 20 items: 19 two-tier multiple-choice questions (with rationale) and one essay question. The test was divided into two comparable packages (A and B). Content validation confirmed that all items met the validity criteria, with Aiken’s V index scores ranging from 0.82 to 0.96, exceeding the cutoff value of 0.75. Validation involved seven experts in science education and learning assessment and evaluation, including two Professors and five Associate Professors with doctoral qualifications, all from four universities in Indonesia. Aligned with the APA Report, the test focused on six core CT skills and their sub-skills, specifically tailored to the science topic of ‘the human circulatory system’ and its sub-topics, as outlined in **Table 2**.

The structure of the selected science topics follows the rules of the national curriculum in Indonesia. Therefore, the field test did not interfere with the learning objectives. The two-tier test allowed students to answer the questions with their conceptual understanding while testing their CT skills based on the stimulus in each item. To measure self-regulation, we use essay questions so that students can reflect on themselves more freely. Based on **Table 1**, there is more than one question on each subskill and subtopic. The aim is to provide sufficient questions in case some questions are invalid and cannot be used. But this does not apply to self-regulation criteria in the form of an open-ended essay. The self-regulation questions for type A and type B have the same meaning but are worded differently. These questions reflect the overall learning outcomes.

An example of a CT item is presented in **Table 3**. The item was used to measure the evaluation skill, the sub-skill of assessing the arguments’ quality. This item is designed to measure students’ ability to determine the quality of arguments based on a given stimulus. It requires students to have a good understanding of the science concepts that support it. The highest score is obtained when students answer the question correctly and provide a correct reason.

**Table 4.** CT skills measurement results

Sub-skill CT	Score 1 (%)	Score 2 (%)	Score 3 (%)	Score 4 (%)
Interpretation	28.88 (13.01)	14.00 (6.31)	7.50 (3.38)	171.63 (77.31)
Analysis	39.33 (17.72)	50.17 (22.60)	16.00 (7.21)	116.50 (52.48)
Evaluation	49.43 (22.27)	40.14 (18.08)	27.29 (12.29)	105.14 (47.36)
Inference	45.50 (20.50)	36.63 (16.50)	11.88 (5.35)	128.00 (57.66)
Explanation	66.22 (29.83)	27.67 (12.46)	21.00 (9.46)	107.11 (48.25)
Self-regulation	128.00 (57.66)	50.50 (22.75)	25.50 (11.49)	18.00 (8.11)

### Validity and Reliability of CT Instruments

The study employed the Rasch model to evaluate the construct validity and reliability, analyzing 445 student responses (including instrument type A and type B). Results from the initial confirmatory factor analysis (CFA) showed strong factor loadings for all dimensions—interpretation, analysis, inference, evaluation, explanation, and self-regulation—ranging from 0.70 to 0.97, indicating robust correlations between items and their latent constructs with minimal measurement error. Construct reliability values ranged from 0.83 to 0.99, demonstrating excellent internal consistency. In the second CFA phase, these CT dimensions maintained high validity and reliability, with factor loadings of 0.83 to 0.98 and construct reliability of 0.97. Combined, the CFA and the Rasch model results confirm the model's robustness and suitability for assessing CT skills in practical applications.

### Statistical Analysis

The analysis of fit statistics in the Rasch model, specifically infit and outfit, aims to determine whether each item within CT aspects aligns with the expectations of the Rasch model. Infit and outfit statistics assess the fit between empirical data and the model's predictions for each item, with ideal values ranging between 0.7 and 1.3 (Bond et al., 2021). Infit or outfit values outside this range indicate misfit, suggesting that an item does not function effectively in measuring the intended ability (Affandy et al., 2021). Misfit values highlight items requiring further examination or adjustment to enhance their accuracy in assessing students' CT. Reliability analysis is also conducted to evaluate the instrument's measurement consistency, encompassing person reliability and item reliability. Person reliability reflects the consistency of students' responses, while item reliability indicates the stability of item difficulty levels. A reliability value exceeding 0.7 is considered satisfactory (Andrich & Marais, 2019), signifying the instrument's reliability for measuring CT skills.

Rating scale model and partial credit model are used to evaluate item difficulty, student ability distribution, and the functionality of graded scales in measurements (Chye & Waugh, 2010). The Rasch model estimates the difficulty parameter for each item within CT domains, such as determining whether items in the evaluation domain are more challenging than those in

interpretation. Additionally, the Rasch model estimates student ability, representing the likelihood of students successfully responding to items (Smith, 2003).

The Wright Map visualizes the relationship between student ability and item difficulty. The Wright Map positions items related to student abilities, enabling the analysis of alignment between item difficulty and student capability. This visual tool is instrumental in assessing whether items are appropriately matched to student abilities or require further adjustment.

Category probability curves (CPC) are employed to assess the effectiveness of graded scales (e.g., 1-5) in differentiating levels of student ability within CT domains. By examining category thresholds, each scale category must exhibit a logical sequence to function effectively in measuring distinct levels of ability. Ordered thresholds indicate that the scale categories offer clear distinctions, with each level progressively representing higher abilities. Conversely, threshold disordering or overlapping—where boundaries between categories are unclear or too close—can compromise the scale's effectiveness (Engelhard & Wang, 2021). CPC also indicate whether the probability of selecting a given category shifts appropriately with increasing ability. If categories show disorder or excessive similarity, category collapse may be considered. We may consider merging similar categories to simplify the scale and improve its ability to differentiate student ability.

## RESEARCH RESULTS

### Measurement of Students' CT Skills in Each Aspect

The measurement results indicate that self-regulation yielded the lowest average CT score among the other sub-skills. The interpretation skill yielded the highest average score within the category of a level 4 rating. The distribution of scores across all aspects indicates that interpretation is the most mastered CT skill among students (see [Table 4](#)). In contrast, self-regulation is the skill that needs the most improvement. A general pattern emerges in which foundational skills such as interpretation and explanation tend to be more developed. In contrast, self-regulation, which requires more reflective thinking skills, remains a significant challenge for most students. This condition highlights the need for targeted instructional strategies to improve students' reflective and self-regulatory skills in CT.

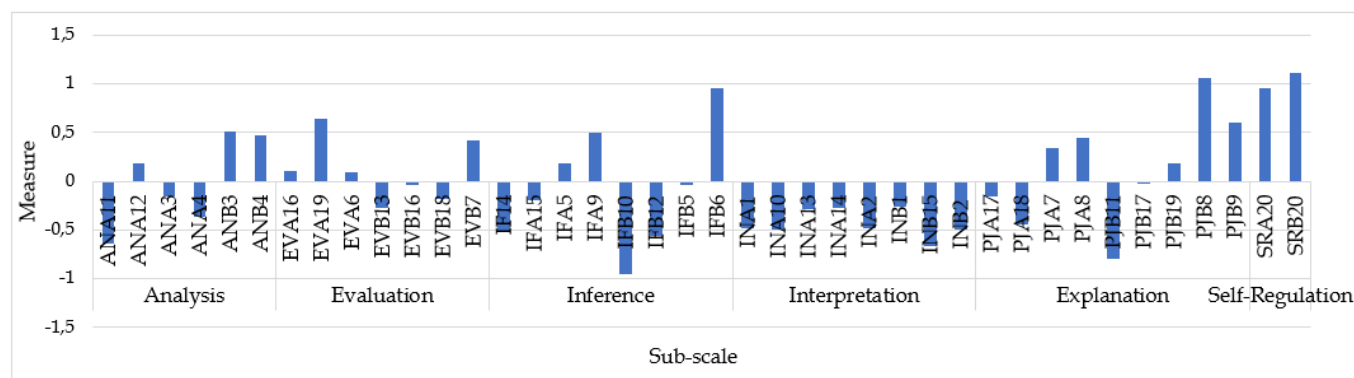


Figure 1. Item difficulty results for each critical thinking aspect (Source: Authors' own elaboration)

The interpretation aspect showed the highest percentage of students scoring in category 4 (77.31%), demonstrating that most students could interpret information well. However, a notable proportion scored in lower categories, particularly in category 1 (28.88%), indicating the need for improvement in basic interpretation skills among some students. The analysis skill revealed a more balanced distribution, with category 4 (52.48%) being the highest, followed by category 2 (22.60%). This result suggests that while most students have adequate analytical abilities, there remains variability that warrants attention for further improvement. In the evaluation skill, scores were relatively evenly spread, with the highest percentages in categories 1 (22.27%) and 4 (47.36%). This condition indicates that while evaluation was challenging for many, certain groups demonstrated strong mastery. Similarly, for the inference, category 4 (57.66%) had the highest percentage, indicating strong understanding among most students, though some scored in lower categories, like 1 (20.50%) and 2 (16.50%), requiring targeted support. The explanation had a significant portion of students in category 1 (29.83%) and category 4 (48.25%), highlighting a mix of basic and advanced skills in this area. However, the self-regulation aspect showed a stark contrast, with most students scoring in category 1 (57.66%)—the highest percentage across all aspects—indicating this skill remains the most difficult. Only a small percentage (8.11%) achieved category 4, emphasizing the need for focused interventions to improve self-regulation.

**Item Difficulty Level For Each Aspect of CT**

The difficulty levels of items across various aspects of CT skills show noticeable variation, reflecting the differences in students' mastery of the assessed sub-skills. These variations are visualized in Figure 1, illustrating item difficulty distribution across CT sub-skills. The analysis aspect showed a difficulty range from -0.64 to 0.51, with item ANA11 (-0.64) being the easiest and ANB3 (0.51) the most challenging. This range indicates varying levels of student mastery for items within the analysis aspect.

For the evaluation aspect, item difficulty ranged from -0.27 (EVB13) to 0.64 (EVA19), with EVB13 being the easiest and EVA19 the hardest. The relatively significant difference in difficulty levels suggests moderate item distribution within this aspect. The inference aspect displayed the widest range, from -0.95 (IFB10) to 0.95 (IFB6). IFB10 was the easiest, while IFB6 was the hardest, highlighting a substantial gap in difficulty levels across items. The interpretation aspect had consistently lower difficulty levels, ranging from -0.66 (INB15) to -0.26 (INB1), with all items scoring negatively. This fact indicates that interpretation is generally easier for students, with no items posing significant difficulty, making it the most accessible CT aspect. The explanation aspect ranged from -0.8 (PJB11) to 1.06 (PJB8). PJB11 was the easiest, while PJB8 was the hardest, reflecting a wide variation in mastery. Some items were challenging, while others were relatively simple. Finally, the self-regulation aspect included only two items with high difficulty levels: SRA20 (0.95) and SRB20 (1.11), the highest values in the dataset. This result indicates that self-regulation is students' most challenging CT skill.

**Category Function of the Polytomous Scale on Each Aspect**

Based on the distribution of scores in each scale category (1-4) for all aspects of CT (sub-skills), it is known that the scale categories did not function fully progressively in distinguishing students' abilities. For example, in the analysis aspect, the distribution of scores in certain categories, such as ANA12, shows a higher frequency in category 2 (91) than in category 1 (28) or category 3 (7). Similar conditions are seen in other items, such as ANB3, which has the highest scores in category 1 (105) and category 4 (81) but lower in category 2 (24) and category 3 (12). This condition shows an inconsistency in distribution, which may indicate that the categories on the scale are not optimal in reflecting the progressive improvement in CT skills. CPC (see Figure 2) are instrumental in assessing the functionality of each category on a polytomous scale, particularly by identifying overlaps or gaps between categories.

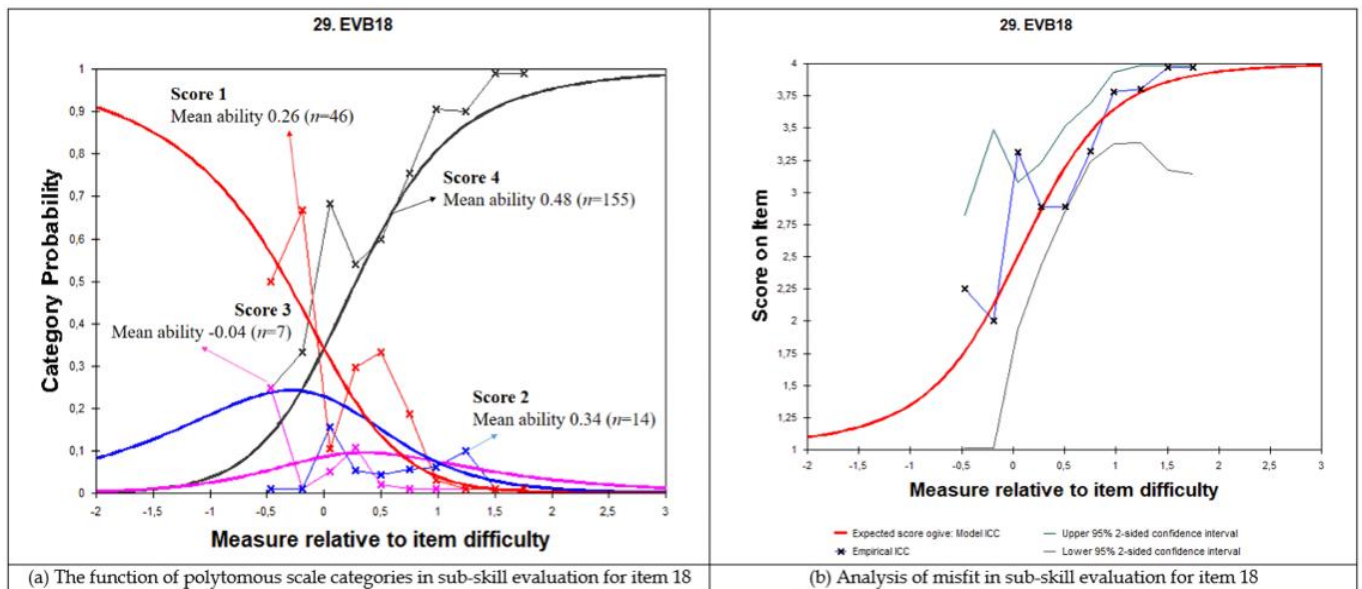


Figure 2. Category function graph for the evaluation skills (Source: Authors' own elaboration)

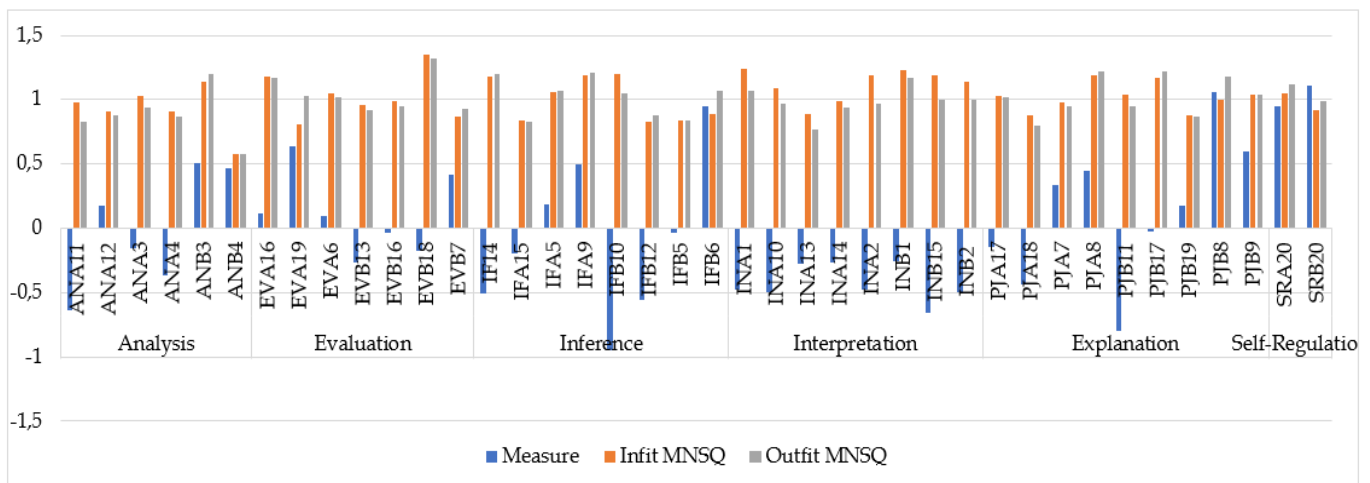


Figure 3. The data of infit and outfit MNSQ for CT (Source: Authors' own elaboration)

In this study, CPC indicated that most test items effectively distinguished between student abilities, reflecting a wedge-like pattern where categories progressively captured different levels of ability. However, two items under the self-regulation aspect (SRA20 and SRB20) failed to differentiate adequately between students.

Both items exhibited concentrated distributions in category 1 and category 2, with steep declines in category 3 and category 4. This pattern suggests insufficient progression across categories, which is likely caused by overlapping thresholds or poorly defined intervals. To address this, the merging or redefinition of categories—such as combining category 2 and category 3 or expanding intervals—may improve the scale's functionality. These adjustments would create more precise distinctions, allowing the scale to capture varying levels of student ability more effectively and ensuring that each category represents a distinct proficiency level.

### Item Misfit on Rasch Analysis

Based on the analysis of infit and outfit mean square (MNSQ) shown in Figure 3, values for CT items, some items demonstrate misfit with the Rasch model, as their values fall outside the acceptable range of 0.6-1.4. In the analysis aspect, item ANB3 has an infit MNSQ value of 1.14 and an outfit MNSQ value of 1.2, which are within the upper boundary but still acceptable. However, item ANB4 shows low compatibility with an infit MNSQ of 0.58, which is below the threshold, suggesting underfitting. In the evaluation aspect, item EVB18 presents higher-than-recommended values, with an infit MNSQ of 1.35 and an outfit MNSQ of 1.32, indicating underfitting with the model. In the inference aspect, items IF14 and IFA9 have infit and outfit values outside the ideal range, with IF14 having infit 1.18 and outfit 1.2, while IFA9 has infit 1.19 and outfit 1.21. Meanwhile, in the interpretation aspect, several items have infit and outfit values greater than 1.4, such as INA1 (infit 1.24 and outfit 1.07) and INB1 (infit 1.23 and outfit 1.17). In

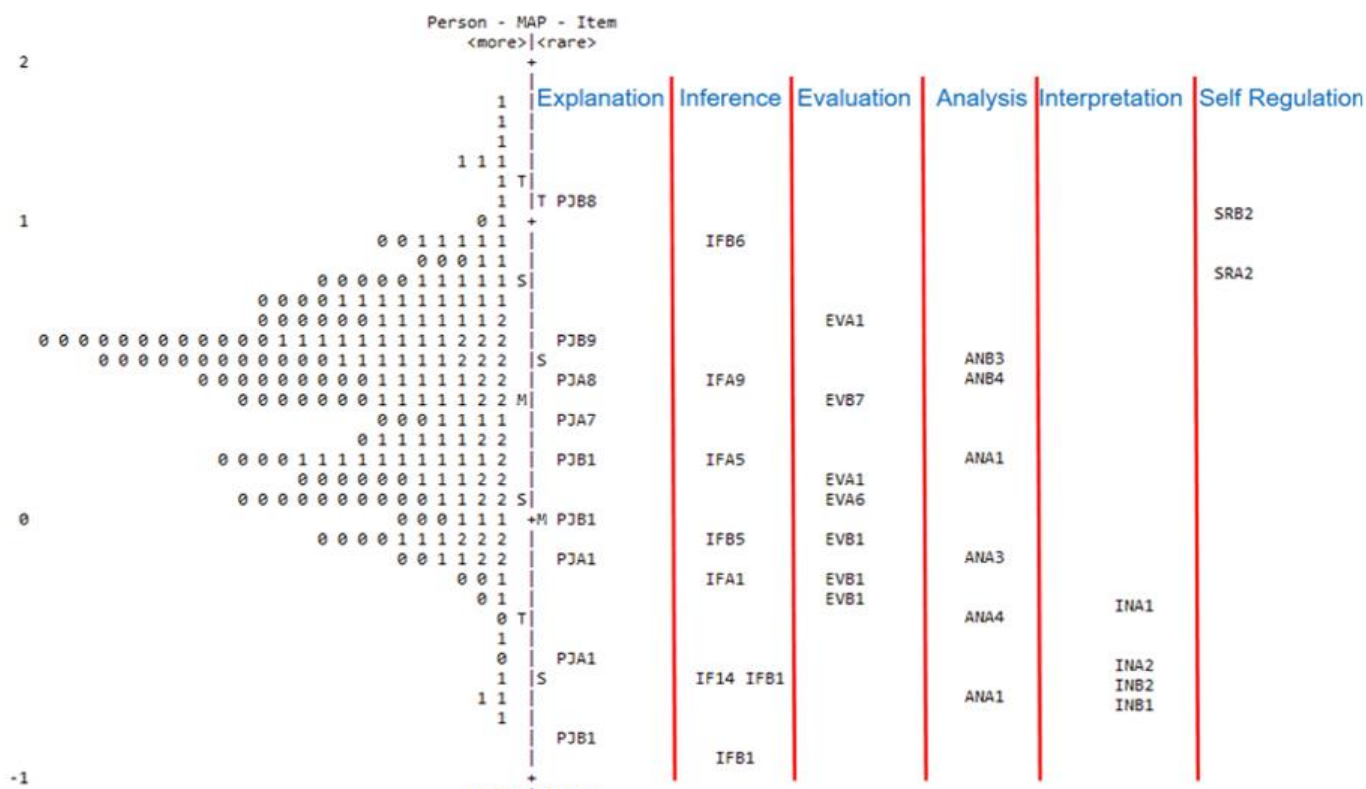


Figure 4. The Wright Map distribution of students' critical thinking skills (Source: Authors' own elaboration)

explanation, items PJA8 and PJB17 also have higher infit and outfit values but are still within the recommended limits.

The distribution of students' CT skills is presented in the Wright Map shown in Figure 4. Based on the map, certain items, such as SRB2 and PJB8, are associated with higher difficulty levels, aligning well with students possessing advanced CT abilities. This result indicates that students with higher skills can tackle more complex challenges, as reflected in these items. However, a noticeable gap exists between student ability levels and item difficulty in certain intervals. Specifically, a significant gap is evident between students with moderate and high abilities, where there is a limited number of items addressing this range. This phenomenon may suggest a shortage of items capable of effectively assessing students at intermediate-to-advanced levels, potentially reducing the measurement tool's effectiveness in identifying CT skills.

## DISCUSSION

The overall level of students' CT is shaped by their performance across six key aspects: interpretation, analysis, evaluation, inference, explanation, and self-regulation. Based on the scores, self-regulation significantly contributed to students' CT, achieving the highest average score (57.66%). This score was followed by explanation (48.25%), inference (57.66%), analysis (52.48%), evaluation (47.36%), and interpretation (77.31%). These results suggest that self-regulation is

dominant, while interpretation scores are comparatively lower.

Facione (2020) emphasizes that these interdependent aspects enhance CT skills. High self-regulation indicates students' ability to manage their thinking processes effectively, bolstering their overall CT skills (Brookhart, 2010). Paul and Elder (2014) argue that CT requires a comprehensive understanding of cognitive aspects like self-regulation and explanation. Robust self-regulation allows students to identify and rectify flaws in their thought processes, while explanation skills demonstrate their ability to elaborate and justify arguments (Sebatana & Dudu, 2022). However, the low score on the interpretation aspect (13.01%) highlights a weakness in understanding and interpreting information, which could impede CT, particularly in analyzing and evaluating arguments.

The self-regulation aspect showed the highest achievement, while the interpretation aspect demonstrated weaker performance. Data indicated that self-regulation scored the highest at 57.66%, compared to the interpretation aspect, with the lowest score of 13.01%. This result suggests that students are more adept at managing their thinking processes independently than accurately interpreting information. Facione (2020) highlights that differences in CT aspect achievement could be influenced by instructional approaches and material complexity. Self-regulation often benefits from explicit classroom practices, such as self-monitoring and reflective activities, which allow students to develop these skills (Gurcay & Ferah, 2018). Zimmerman &



Schunk (2004) also emphasize the role of structured training in fostering self-regulation through guided evaluation and control of thought processes.

Meanwhile, low interpretation scores might stem from insufficient practice or challenges associated with complex content requiring deeper analysis (Wang et al., 2017). However, high scores in self-regulation do not always reflect genuine mastery. Some students may perform well due to external factors, such as environmental expectations or habitual learning behaviors, rather than a thorough understanding of CT strategies. Additionally, low interest in particular topics may contribute to weaker performance in interpretation tasks.

The learning strategies should emphasize improving CT aspects with lower scores, particularly interpretation, to enhance overall student CT. The interpretation aspect scored the lowest at 13.01%, highlighting a significant weakness in students' ability to understand and make sense of the information presented. Strong interpretation skills are essential for contextual understanding, identifying critical information, and developing a comprehensive grasp of the subject matter (Chudgar et al., 2016; Lukman et al., 2021). Weak interpretation abilities can impede CT skills, such as analysis and evaluation, which depend on a solid foundation of understanding the initial information (Danday, 2021). Research by Brookhart (2010) suggests that interpretation skills can be strengthened through instructional strategies like structured discussions, concept mapping, and text analysis, all of which help students process and contextualize information.

Moreover, project-based learning and inquiry-based questioning strategies have effectively encouraged students to interpret data and information actively (Jansen & Söbke, 2022; Khaeruddin et al., 2023). However, while focusing on interpretation, it is crucial to maintain a balanced approach supporting the development of other CT aspects, such as self-regulation and evaluation. Overemphasis on one aspect may limit students' ability to integrate and apply CT skills across diverse contexts synergistically.

The difficulty levels across various sub-skills of CT showed diverse results, with specific items being identified as the easiest or most challenging. Higher values indicate that respondents found the items more accessible, whereas lower values signify more incredible difficulty (Bond et al., 2021). For the analysis aspect, item ANB3 had the highest difficulty value (0.51), making it the easiest, while ANA11 had the lowest value (-0.64), indicating it was the most difficult. In the evaluation aspect, EVA19 was the easiest (0.64), whereas EVB13 was the most challenging (-0.27). For inference, the easiest item was IFB6 (0.95), and the hardest was IFB10 (-0.95). In the interpretation aspect, the easiest was INB1 (-0.26), and the hardest was INB15 (-0.66).

The explanation aspect had PJB8 as the easiest (1.06) and PJB11 as the hardest (-0.8). Lastly, in self-regulation sub-skill, SRB20 had the highest value (1.11), and notably, no items had negative values, indicating a general tendency of difficulty for this aspect.

The inference aspect demonstrates a higher difficulty level than others, as reflected in the relatively more extensive and extreme distribution of negative scores in the inference subscale. It indicates that students generally experience more difficulty answering items within this subscale. Based on the research findings, several items in the subscale of inference have negative values, including IF14 (-0.51), IFB10 (-0.95), IFB12 (-0.56), and IFB5 (-0.04). These negative scores highlight that this aspect is more complex than others, such as self-regulation, which has entirely positive scores, or evaluation, which exhibits a more balanced distribution of positive and negative scores.

Theoretically, inference is a high-level cognitive process requiring combining information from multiple sources and a deep understanding to draw valid conclusions (Facione, 2000). Cognitive psychology research indicates that inference involves the ability to connect abstract information, often posing challenges for students, particularly if they lack sufficient background knowledge on the topic (Townend & Brown, 2016).

The Wright Map data illustrates the distribution of items on a logit scale ranging from -1 to +2. Higher logits represent higher student ability, while lower logits reflect lower student ability. For example, item "SRB2," located at logit +1, tends to challenge students with high ability, whereas item "ANA4," positioned at logit 0, is more suitable for students with average ability. The positioning of items on the Wright Map reflects their difficulty level relative to student ability (Smith, 2003). Students with abilities equal to or exceeding the item's difficulty level are more likely to answer correctly (Engelhard & Wang, 2021). Conversely, items with a difficulty level above a student's ability are generally answered incorrectly (Smith, 2003).

The Wright Map data indicates that the distribution of items effectively covers a wide range of student abilities. For instance, items at logit +1 (such as "SRB2") challenge high-ability students due to their positioning above average students' ability. In contrast, items between logits -1 and 0 (such as "INA1" and "ANA4") are more accessible to lower-ability students as their difficulty level aligns with or is below the average ability. This distribution demonstrates that the items in each CT aspect are sufficiently varied to address students with different ability levels. Items located at high logits are best suited for students with advanced CT skills, while lower logits are more appropriate for those with lower CT skills.

## CONCLUSIONS AND IMPLICATIONS

The findings of this study reveal that self-regulation received the lowest average score among students, indicating a significant area for improvement. The students may struggle with managing their cognitive and metacognitive processes when engaging with complex problems in the human circulatory system. In contrast, interpretation emerged as the most mastered skill, particularly in the level 4 rating category. This result aligns with the findings of Tiruneh (2014) and Abrami (2015) in their meta-analysis studies, which highlighted that students tend to master interpretation more readily, as these skills are closely tied to the fundamental process of understanding and engaging with provided information. Educators should involve various learning innovations, such as problem-based learning, inquiry-based learning, and STEM-based learning, which have been shown to enhance CT skills (Sujatmika et al., 2024).

However, this study has limitations related to its specific geography, which may affect the generalizability of the findings to other regions with different socioeconomic and cultural contexts. Future research could explore the interplay between CT and other factors, such as gender and school location, particularly in relation to the availability of resources and facilities that support effective science learning. Investigating these aspects would provide a more comprehensive understanding of the influences on developing CT. Additionally, longitudinal studies could be conducted to examine how CT skills evolve over time and across varying educational environments.

**Author contributions:** SS: conceptualization, methodology, formal analysis, investigation, writing - original draft; S: supervision, validation, writing - review & editing; MM: supervision, methodology, validation, writing - review & editing; BAP: supervision, methodology, validation, writing - review & editing. All authors agreed with the results and conclusions.

**Funding:** No funding source is reported for this study.

**Ethical statement:** The authors stated that the study has received approval from Universitas Sebelas Maret under letter number 17960/UN27/PK.03.08/2024. Informed consent was obtained from all participants after they were provided with a detailed explanation of the study's objectives, procedures, and potential risks. Participation was entirely voluntary, with no coercion involved at any stage. The authors further stated that all personal data, including names, gender, schools, class, and age, were anonymized and securely stored. Access to these data was strictly limited to authorized researchers to ensure confidentiality and compliance with ethical research guidelines.

**Declaration of interest:** No conflict of interest is declared by the authors.

**Data sharing statement:** Data supporting the findings and conclusions are available upon request from the corresponding author.

## REFERENCES

- Abosalem, Y. (2015). Assessment techniques and students' higher-order thinking skills. *International Journal of Secondary Education*, 4(1), 61-66. <https://doi.org/10.11648/j.ijsedu.20160401.11>
- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85(2), 275-314. <https://doi.org/10.3102/0034654314551063>
- Affandy, H., Nugraha, D. A., Pratiwi, S. N., & Cari, C. (2021). Calibration for instrument argumentation skills on the subject of fluid statics using item response theory. *Journal of Physics: Conference Series*, 1842, Article 012032. <https://doi.org/10.1088/1742-6596/1842/1/012032>
- Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer. <https://doi.org/10.1007/978-981-13-7496-8>
- Bailin, S. (2002). Critical thinking and science education. *Science & Education*, 11, 361-375. <https://doi.org/10.1023/A:1016042608621>
- Bao, L., & Koenig, K. (2019). Physics education research for 21st-century learning. *Disciplinary and Interdisciplinary Science Education Research*, 1, Article 2. <https://doi.org/10.1186/s43031-019-0007-8>
- Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. <https://doi.org/10.4324/9780429030499>
- Brookfield, S. D. (2017). *Becoming a critically reflective teacher*. John Wiley & Sons.
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.
- Castro, R. (2019). Blended learning in higher education: Trends and capabilities. *Education and Information Technologies*, 24(4), 2523-2546. <https://doi.org/10.1007/s10639-019-09886-3>
- Chudgar, S. M., Engle, D. L., Grochowski, C. O. C., & Gagliardi, J. P. (2016). Teaching crucial skills: An electrocardiogram teaching module for medical students. *Journal of Electrocardiology*, 49(4), 490-495. <https://doi.org/10.1016/j.jelectrocard.2016.03.021>
- Chye, T. H., & Waugh, R. F. (2010). A unidimensional Rasch measure of motivation in science and mathematics. In R. Waugh (Ed.), *Applications of Rasch measurement in education* (pp. 77-94). Nova Science Publishers.
- Danday, B. A. (2021). Advancing preservice physics teachers' critical thinking through active and passive microteaching lesson study. *International Journal of Learning, Teaching and Educational Research*, 20(3), 205-228. <https://doi.org/10.26803/ijlter.20.3.13>

- Davies, M. (2015). A model of critical thinking in higher education. In M. Paulsen (Ed.), *Higher education: Handbook of theory and research, vol 30* (pp. 41-92). Springer. [https://doi.org/10.1007/978-3-319-12835-1\\_2](https://doi.org/10.1007/978-3-319-12835-1_2)
- Dwyer, C. P., Hogan, M. J., Harney, O. M., & Kavanagh, C. (2017). Facilitating a student-educator conceptual model of dispositions towards critical thinking through interactive management. *Educational Technology Research and Development*, 65(1), 47-73. <https://doi.org/10.1007/s11423-016-9460-7>
- Dwyer, C. P., Hogan, M. J., Harney, O. M., & O'Reilly, J. (2014). Using interactive management to facilitate a student-centred conceptualization of critical thinking: A case study. *Educational Technology Research and Development*, 62(6), 687-709. <https://doi.org/10.1007/s11423-014-9360-7>
- Engelhard, G., & Wang, J. (2021). *Rasch models for solving measurement problems: Invariant measurement in the social sciences*. SAGE. <https://doi.org/10.4135/9781071878675>
- Ennis, R. (1991). Critical thinking. *Teaching Philosophy*, 14(1), 5-24. <https://doi.org/10.5840/teachphil19911412>
- Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal Logic*, 20(1), 61-84. <https://doi.org/10.22329/il.v20i1.2254>
- Facione, P. A. (2020). Critical thinking: What it is and why it counts. *Insight Assessment*. <https://www.law.uh.edu/blakely/advocacy-survey/Critical%20Thinking%20Skills.pdf>
- Forawi, S. A. (2016). Standard-based science education and critical thinking. *Thinking Skills and Creativity*, 20, 52-62. <https://doi.org/10.1016/j.tsc.2016.02.005>
- Franco, A., Marques Vieira, R., & Tenreiro-Vieira, C. (2018). Educating for critical thinking in university: The criticality of critical thinking in education and everyday life. *Journal for Communication Studies*, 11(2), 131-144.
- García-Carmona, A. (2023). Scientific thinking and critical thinking in science education: Two distinct but symbiotically related intellectual processes. *Science and Education*, 34, 227-245. <https://doi.org/10.1007/s11191-023-00460-5>
- Gurcay, D., & Ferah, H. (2018). High school students' critical thinking related to their metacognitive self-regulation and physics self-efficacy beliefs. *Journal of Education and Training Studies*, 6(4), 125-130. <https://doi.org/10.11114/jets.v6i4.2980>
- Hakim, N. W. A., & Talib, C. A. (2018). Measuring critical thinking in science: Systematic review. *Asian Social Science*, 14(11), 9-15. <https://doi.org/10.5539/ass.v14n11p9>
- Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, 53(4), 449-455. <https://doi.org/10.1037/0003-066X.53.4.449>
- Howard, L. W., Tang, T. L. P., & Austin, M. J. (2015). Teaching critical thinking skills: Ability, motivation, intervention, and the Pygmalion effect. *Journal of Business Ethics*, 128(1), 133-147. <https://doi.org/10.1007/s10551-014-2084-0>
- Jansen, E., & Söbke, H. (2022). Communication skills in construction projects and promoting them through multiplayer online games. In H. Söbke, P. Spangenberg, P. Müller, & S. Göbel (Eds.), *Serious games. JCSG 2022. Lecture notes in computer science, vol 13476* (pp. 169-181). Springer. [https://doi.org/10.1007/978-3-031-15325-9\\_13](https://doi.org/10.1007/978-3-031-15325-9_13)
- Khaeruddin, K., Indarwati, S., Sukmawati, S., Hasriana, H., & Afifah, F. (2023). An analysis of students' higher-order thinking skills through the project-based learning model on science subject. *Jurnal Pendidikan Fisika Indonesia*, 19(1), 47-54. <https://doi.org/10.15294/jpfi.v19i1.34259>
- Kuhn, D. (2019). Critical thinking as discourse. *Human Development*, 62(3), 146-164. <https://doi.org/10.1159/000500171>
- Lai, E. R. (2011). Critical thinking: A literature review. *Pearson Research Reports*, 6(1), 40-41.
- Liu, Y., & Pásztor, A. (2022). Effects of problem-based learning instructional intervention on critical thinking in higher education: A meta-analysis. *Thinking Skills and Creativity*, 45, Article 101069. <https://doi.org/10.1016/j.tsc.2022.101069>
- Lombard, F., Schneider, D. K., Merminod, M., & Weiss, L. (2020). Balancing emotion and reason to develop critical thinking about popularized neurosciences. *Science and Education*, 29(5), 1139-1176. <https://doi.org/10.1007/s11191-020-00154-2>
- Lukman, Marsigit, Istiyono, E., Kartowagiran, B., Retnawati, H., Kistoro, H. C. A., & Putranta, H. (2021). Effective teachers' personality in strengthening character education. *International Journal of Evaluation and Research in Education*, 10(2), 512-521. <https://doi.org/10.11591/ijere.v10i2.21629>
- Ma, X., Zhang, Y., & Luo, X. (2023). Students' and teachers' critical thinking in science education: Are they related to each other and with physics achievement? *Research in Science and Technological Education*, 41(2), 734-758. <https://doi.org/10.1080/02635143.2021.1944078>
- Mapeala, R., & Siew, N. M. (2015). The development and validation of a test of science critical thinking for

- fifth graders. *SpringerPlus*, 4, Article 741. <https://doi.org/10.1186/s40064-015-1535-0>
- Marin, L. M., & Halpern, D. F. (2011). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity*, 6(1), 1-13. <https://doi.org/10.1016/j.tsc.2010.08.002>
- OECD. (2019). What is PISA? OECD. <http://www.oecd.org/pisa/>
- Paul, R., & Elder, L. (1992). Critical thinking: What, why, and how. *New Directions for Community Colleges*, 77(2), 3-24. <https://doi.org/10.1002/cc.36819927703>
- Paul, R., & Elder, L. (2014). *Consequential validity: Using assessment to drive instruction*. Foundation for Critical Thinking.
- Santos, L. F. (2017). The role of critical thinking in science education. *Journal of Education and Practice*, 8(20), 159-173.
- Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, 39, 37-63. <https://doi.org/10.1016/j.dr.2015.12.001>
- Sebatana, M. J., & Dudu, W. T. (2022). Reality or mirage: Enhancing 21st-century skills through problem-based learning while teaching particulate nature of matter. *International Journal of Science and Mathematics Education*, 20, 963-980. <https://doi.org/10.1007/s10763-021-10206-w>
- Smith, R. M. (2003). *Rasch measurement models: Interpreting WINSTEPS/BIGSTEPS and FACETS output*. JAM Press.
- Sujatmika, S., Masykuri, M., Prayitno, B. A., & Sutarno, S. (2024). Fostering critical thinking in science education: Exploring effective pedagogical models. *International Journal of Advanced and Applied Sciences*, 11(7), 149-159. <https://doi.org/10.21833/ijaas.2024.07.016>
- Sujatmika, S., Widyawati, A., Ernawati, T., & Widhy, P. (2022). Gambier's product (*Uncaria gambir* Roxb.) as learning material to enhance critical thinking. *AIP Conference Proceedings*, 2600(1), Article 050003. <https://doi.org/10.1063/5.0112224>
- Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. *International Journal of Science and Mathematics Education*, 15(4), 663-682. <https://doi.org/10.1007/s10763-016-9723-0>
- Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies*, 4(1), 1-17. <https://doi.org/10.5539/hes.v4n1p1>
- Townend, G., & Brown, R. (2016). Exploring a sociocultural approach to understanding academic self-concept in twice-exceptional students. *International Journal of Educational Research*, 80, 15-24. <https://doi.org/10.1016/j.ijer.2016.07.006>
- Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. Jossey-Bass.
- United Nations. (2018). Revision of world urbanization prospects. *Department of Economic and Social Affairs, Population Division*. <https://population.un.org/wup/>
- van Laar, E., van Deursen, A. J. A. M., van Dijk, J. A. G. M., & de Haan, J. (2020). Determinants of 21st-century skills and 21st-century digital skills for workers: A systematic literature review. *SAGE Open*, 10(1). <https://doi.org/10.1177/2158244019900176>
- Wang, H. H., Chen, H. T., Lin, H. S., Huang, Y. N., & Hong, Z. R. (2017). Longitudinal study of a cooperation-driven, socio-scientific issue intervention on promoting students' critical thinking and self-regulation in learning science. *International Journal of Science Education*, 39(15), 2002-2026. <https://doi.org/10.1080/09500693.2017.1357087>
- Zenker, F. (2018). Introduction: Reasoning, argumentation, and critical thinking instruction. *Topoi*, 37(1), 91-92. <https://doi.org/10.1007/s11245-016-9416-x>
- Zimmerman, B. J., & Schunk, D. H. (2004). Self-regulating intellectual processes and outcomes: A social cognitive perspective. In D. Y. Dai, & R. J. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (pp. 523-549). Erlbaum Associates.

<https://www.ejmste.com>