Volume 4, Issue 3 (2026) 317-333

EQUATOR SCIENCE JOURNAL (ESJ)
e-ISSN: 2988-3881
Journal homepage:
https://ejournal.edukhatulistiwa.com/index.php/ESJ

Content Validation of a Critical Thinking Skills Test on Temperature and Heat (CTS-THT)

Sunaryo Romli[1], Andi Suhandi[1*], Muslim Muslim[1], Ida Kaniawati[1], Mohd Zaidi Bin Amiruddin[1]
[1]Science Education Study Program, Universitas Pendidikan Indonesia, Indonesia

Received: May 31, 2026 | Revised: July 21, 2026 | Accepted: August 5, 2026
First Available Online: August 11, 2026 | Publication Date: October 31, 2026
DOI: https://doi.org/10.61142/esj.v4i3.405

*Corresponding author: andi_sh@upi.edu

Keywords:
Critical thinking;
Science education;
Secondary student;
Temperature and heat;
Validation

ABSTRACT
This study aims to develop and validate a Critical Thinking Skills Test on Temperature and Heat (CTS-THT) for junior high school students based on Ennis's critical thinking framework. The research employs an instrument development design, beginning with a literature review, followed by a test blueprint operationalizing five aspects of critical thinking: basic clarification, basic support, inference, advanced clarification, and strategies and tactics into essay items. The CTS-THT assesses students' ability to analyze arguments, interpret experimental data, define key physics terms, and justify decisions related to temperature and heat phenomena. Content validity is examined through expert judgment by five validators: two experts in physics education, two in educational assessment, and one in science content. The validators rate each item on five dimensions using a 0–1 scale, and content validity is quantified using Aiken's V index. The results show that all 42 items achieve Aiken's V coefficients of at least 0.60, indicating satisfactory content validity. Most items are valid without revision, while only a few require minor or substantive improvements. These findings indicate that the CTS-THT is a content-valid instrument for assessing junior high school students' critical thinking skills on temperature and heat. Theoretically, this study contributes a domain-specific, Ennis-based instrument addressing gaps in existing critical thinking and physics tests. Practically, the CTS-THT provides a curriculum-aligned instrument to help science teachers assess and foster students' critical thinking through more targeted instruction and assessment.

INTRODUCTION

Critical thinking skills are a central goal in science education, particularly in physics, because students must not only understand scientific concepts but also interpret information, analyze causal relationships, evaluate evidence, and draw rational conclusions when solving problems (Popescu & Morgan, 2007; Parkkinen et al., 2018; Wei et al., 2022; Saputri et al, 2026). In physics learning, these skills are crucial for helping students understand more complex scientific phenomena and develop appropriate solutions to the problems they face in real-world contexts. Concepts such as temperature, heat, and thermal energy are fundamental to thermodynamics but often lead to misconceptions among students, so they require learning and assessment approaches that explicitly foster and assess critical thinking (Velmovská & Bartošovič, 2016; Seibert, 2021; Gavronskaya et al., 2022; Pramasdyahsari et al., 2023; Rahayu & Alsulami, 2024).

Ennis (1985) defines critical thinking as rational, reflective thought focused on deciding what to believe or do. Based on this conception, Ennis proposes a framework comprising several core abilities that can be operationalized in assessment instruments. In this study, five main aspects of Ennis's framework serve as the conceptual basis for test development: providing basic clarification, building basic support, making inferences, making advanced clarifications, and employing strategies and tactics (Ennis, 1985). Basic clarification involves clarifying questions and analyzing existing arguments; in the context of temperature and heat, this entails understanding the meaning of physics questions and identifying key information in physical situations. Basic support refers to assessing the credibility of information sources and interpreting observational data, as reflected here in students' skills in evaluating data and information from temperature and heat experiments. The inference aspect involves drawing logical conclusions or interpretations from available evidence; in physics learning, it is closely related to interpreting experimental results and generalizing about temperature and heat processes. Advanced clarification emphasizes defining terms and identifying underlying assumptions in an argument, which is particularly important when students distinguish among temperature, heat, and thermal energy and when they connect macroscopic observations with microscopic explanations. The strategies and tactics aspect concerns choosing appropriate courses of action and interacting with others to solve problems or make rational decisions. In this study, these abilities are assessed by students' capacity to choose and justify solution strategies for physics problems involving temperature and heat in realistic contexts.

Taken together, these five aspects form an integrated theoretical framework that directly links critical thinking theory to physics assessment of temperature and heat. In this framework, each Ennis (1985) ability is aligned with a specific physics-reasoning competency and a corresponding item demand: basic clarification maps to interpreting the meaning of thermal problems and identifying relevant physical quantities; basic support maps to evaluating experimental data on temperature and heat; inference maps to drawing conclusions from calorimetric or thermal-equilibrium evidence; advanced clarification maps to defining and distinguishing temperature, heat, and thermal energy at the macroscopic and microscopic levels; and strategies and tactics map to selecting and justifying solutions to authentic thermal problems. This alignment provides the conceptual bridge between the critical thinking construct and the content domain and serves as the blueprint for item writing, expert judgment, and the interpretation of validity evidence in the present study.

Measuring critical thinking skills in science education is important to ensure that students not only understand physics concepts theoretically but also apply this knowledge to reasoning, evaluating arguments, and complex problem-solving. Previous studies have highlighted the importance of using standardized, validated instruments to assess critical thinking skills in science education, so that assessment results truly reflect the competencies educators intend to develop (Ennis, 1993). For many years, general standardized instruments such as the *Watson-Glaser Critical Thinking Appraisal* and the *Halpern Critical Thinking Assessment* (HCTA) have been widely used to measure critical thinking skills in a broad sense (Bernard et al., 2008; Butler, 2012; Ghanizadeh et al., 2024). However, because these instruments are not specific to physics, they are less accurate at measuring the critical thinking required to understand and apply physics concepts in areas such as temperature and heat.

Recent years have seen increased focus on developing domain-specific tools to assess critical thinking in science education. For example, Sartono et al. (2025) designed an instrument to measure students' critical thinking skills in basic thermodynamics and heat transfer, including topics such as temperature and heat transfer. They validated the content using Aiken's V index to ensure that each test item aligned with learning objectives and students' cognitive levels. Results showed that the instrument was both valid and reliable for evaluating critical thinking in physics. At the same time, physics education research emphasizes the importance of rigorous validation methods and detailed item analysis to ensure that school assessments accurately and effectively measure students' critical thinking within specific science disciplines.

Recent trends also show growing interest in developing context-based critical thinking assessment instruments in science that align with curriculum demands and comprehensively capture students' analytical, evaluative, inferential, and problem-solving abilities. At the same time, pedagogical approaches such as inquiry-based learning (IBL) and problem-based learning (PBL) are effective in developing students' critical thinking skills by engaging them in investigating problems, analyzing evidence, and making data-driven decisions (Dewi et al., 2025). This suggests that assessment instruments for critical thinking in physics should be compatible with these approaches and capable of capturing higher-order thinking in authentic learning tasks. However, despite various developments in instrument design and validation, instruments that specifically measure junior high school students' critical thinking skills on temperature and heat, explicitly based on Ennis's framework and validated through expert judgment using indices such as Aiken's V, remain limited (Sartono et al., 2025).

This situation highlights a significant research gap: existing tools are either broad critical-thinking assessments not specific to physics or physics tests that do not fully capture critical-thinking indicators for topics such as temperature and heat. Consequently, there's a need for instruments grounded in a strong critical-thinking framework, tailored to the junior high school physics curriculum on temperature and heat, and rigorously validated through content validity analysis. Addressing this gap is essential to ensure that physics critical-thinking assessments reliably measure students' understanding and application of scientific concepts, thereby enhancing teaching and evaluation practices. Unlike Sartono et al. (2025), who developed a critical-thinking assessment on temperature and heat for senior secondary physics, and Illene et al. (2023), who focused on multiple-choice items assessing 21st-century skills in heat and heat transfer, the present study targets junior high school students, operationalizes all five Ennis (1985) aspects into open-ended essay items, and validates them item-by-item using Aiken's V. This combination of an explicit Ennis framework, a junior high school temperature-and-heat context, and per-item content-validity indexing differentiates the CTS-THT from previously developed instruments. Methodologically, the study contributes item-level content-validity evidence: rather than reporting a single instrument-level index, each of the 42 items is quantified with Aiken's V across five judgment dimensions, yielding a transparent, item-by-item basis for retention or revision that is rarely reported for junior high school critical-thinking instruments.

Therefore, this study aims to develop and validate a Critical Thinking Skills Test on Temperature and Heat (CTS-THT) for junior high school students, grounded in Ennis's (1985) critical thinking framework and aligned with the temperature and heat content of the science curriculum. Specifically, the objectives of this research are: (1) to construct a set of test items that operationalize the five aspects of Ennis's critical thinking framework (basic clarification, basic support, inference, advanced clarification, and strategies and tactics) in the context of temperature and heat; and (2) to evaluate the content validity of the CTS-THT items through expert judgment using Aiken's V index (Aiken, 1985).

METHOD

The instrument development in this study follows a content-oriented test-development model, operationalized as a modified version of the Wesnawa et al. (2022) procedure, integrated with the scaffolding strategy of Nurhayati et al. (2026). This model belongs to the define-design-develop family of development research; in the present study, it is deliberately bounded to the design and content-validation stages, while empirical tryout and psychometric analysis are reserved for subsequent research.

Concretely, the developed instrument comprises four elements: (a) a test blueprint (grid) that maps each Ennis aspect to operational indicators and to temperature-and-heat subtopics; (b) a set of 42 open-ended essay items; (c) an analytic 0-3 scoring rubric; and (d) a five-dimensional expert-validation sheet. Making these elements explicit clarifies the study's developmental contribution. Methodologically, content validation is treated here as the first stage of a broader, multi-stage validation framework: Stage 1 (reported in this article) establishes content validity through expert judgment, quantified with Aiken's V; Stage 2 will establish empirical evidence through a student field trial, including reliability estimation (e.g., Cronbach's alpha and Rasch person–item reliability); and Stage 3 will examine item difficulty, discrimination, and construct validity using Item Response Theory (Rasch modeling) and factor-analytic procedures. This staged design delimits the scope of the present study while making the pathway to full psychometric validation explicit.

This research involves developing an assessment instrument in an essay-only format. The instrument's development follows the guidelines by Wesnawa et al. (2022), which include six steps: defining variables, elaborating indicators, preparing items, trialing the instrument, analyzing its content, and finalizing it. Additionally, Nurhayati et al. (2026) provide a scaffolding strategy for critiquing questions in five steps: describing critical-thinking aspects, operationalizing these aspects, creating a question framework, writing items, and developing a scoring rubric. By integrating these approaches, this study describes critical thinking skills, operationalizes them, creates a question framework, writes questions, develops a scoring rubric, and validates the instrument with experts, who revise it based on their feedback. The instrument is named the Critical Thinking Skills Test on Temperature and Heat (CTS-THT). Its development steps are illustrated in Figure 1. The process began with a review of the literature on critical thinking in science education, the measurement of critical thinking using Ennis's (1985) framework, and an examination of temperature and heat concepts at the junior high level. An instrument grid was then created, outlining critical thinking aspects, operationalizing them, providing question scenarios, and mapping relationships among the indicators, temperature and heat subtopics, and question types.

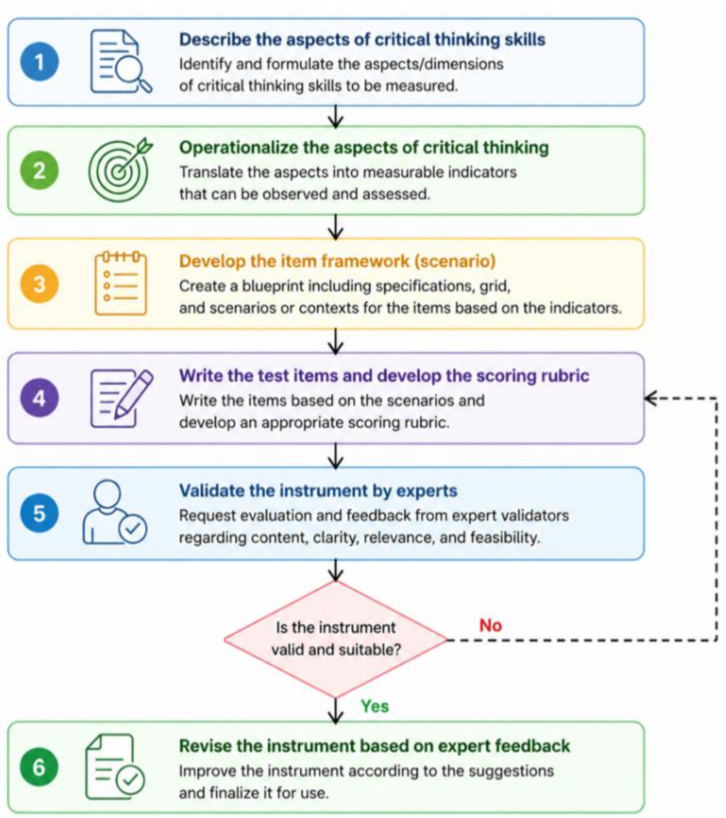

Fig.1. Steps to develop a critical thinking skills test instrument on temperature and heat content (CTS-THT) for secondary students
(adapted from Wesnawa et al. (2022) and Nurhayati et al. (2026))

Based on the grid, the questions are arranged as essays. Each item is designed to assess junior high school students' ability to understand and apply the concepts of temperature and heat, with a focus on analysis, evaluation, and inference. Based on Ennis's (1985) critical thinking framework, this instrument includes five main aspects: providing basic clarification, building basic support, making inferences, making advanced clarifications, and employing strategies and tactics. This instrument consists of essay questions designed to test students' ability to analyze arguments, assess the credibility of sources, make evidence-based decisions, and understand physics concepts in depth.

The assessment rubric uses a 0-3 scale to measure the quality of students' answers, particularly the accuracy and depth of their argumentation and the basis of assessment. This scoring rubric assesses students' answers based on the accuracy of the response and the quality of the explanation provided. A score of 3 is given if both the answer and explanation are correct, indicating a proper understanding. A score of 2 is awarded if the answer is correct but the explanation is somewhat inaccurate. A score of 1 is awarded if the answer is correct but the explanation is incorrect. A score of 0 is assigned in various conditions: if both the answer and explanation are wrong, if the answer is wrong even though the explanation is correct, or if both the answer and explanation are inaccurate. This rubric emphasizes the connection between correct answers and accurate explanations to comprehensively evaluate students' understanding.

The prepared instruments are then given to five expert validators for assessment using validation sheets. Assessments are conducted for all question items across five validation aspects: Ambiguity of Sentence Grammar (ASG), Ease of Understanding Question Intent (EUQ), Suitability of Breadth and Depth of Content (SBC), Accuracy of the Answer Key (AAK), and Conformity of the Question with the Critical Thinking Indicator (CQC). At this stage, the validators also provide qualitative input on concept accuracy, editorial clarity, the potential for multiple interpretations, and the suitability of the questions for critical-thinking indicators and for the characteristics of junior high school students.

Each question item was evaluated against the identified indicators of critical thinking skills, and validator feedback included suggestions to improve conceptual clarity, question wording, and alignment with these indicators. The suggested follow-up for each question includes the revisions needed to make this instrument measure students' critical thinking skills more effectively.

Question items that receive a low Aiken's V score or substantial revision comments are revised based on the validator's input. Revisions address editorial aspects, the question's context, the depth of the concept, the answer key, and alignment with the critical thinking indicators. For example, items considered ambiguous or not conceptually correct, such as questions 2, 17, and 38, are revised to be more accurate, less open to multiple interpretations, and better aligned with the purpose of measurement.

Participant and Data Collection

The participants in this study are five validators who were purposively selected based on academic competence and professional experience relevant to the development of science/physics learning instruments. The five validators consist of two experts in physics education, two experts in the evaluation of educational instruments, and one expert in science/physics content. Validators were selected based on the following criteria: having a minimum of a master's degree in a relevant field; having experience in physics education, educational assessment, or instrument development; understanding temperature and heat materials at the junior high school level; and having at least 10 years of experience teaching, researching, or studying learning evaluation instruments.

Two physics education experts assess the material's suitability, the depth of the concepts, and the relevance of the questions to junior high school physics learning. Two instrument evaluation experts assess the quality of the item constructs, the clarity of the statements, and the instruments' alignment with the principles of educational measurement. One science/physics content expert helps strengthen the assessment of scientific truth, concept accuracy, and answer accuracy. At this stage, junior high school students have not been included as respondents because the research focuses on validating the instrument's content.

Two types of instruments were used in this study. First, the primary instrument was the Critical Thinking Skills Test (CTS-THT) for temperature and heat materials. This instrument consisted of essay questions designed to measure indicators of critical thinking skills, namely analysis, evaluation, inference, and reasoning. We developed the essay items by integrating critical thinking indicators with temperature- and heat-related content indicators relevant to junior high school students.

Second, the validator assessment instrument is an expert validation sheet used to assess each question item across five main aspects. The aspects, descriptions, and scoring criteria are presented in Table 1.

Table 1. Expert Validation Aspects and Scoring Criteria

Aspect	Description	Scoring
Ambiguity of Sentence Grammar (ASG)	Clarity of the language and phrasing of the question.	1 = Unambiguous; 0 = Ambiguous
Ease of Understanding Question Intent (EUQ)	Ease with which students understand the purpose of the question.	1 = Easy; 0 = Not easy
Suitability of Breadth and Depth of Content (SBC)	Suitability of the material coverage with the topic of temperature and heat.	1 = Appropriate; 0 = Not appropriate
Accuracy of the Answer Key (AAK)	Accuracy of the answer based on the correct physics concept.	1 = Correct; 0 = Incorrect
Conformity of the Question with the Critical Thinking Indicator (CQC)	Extent to which the item aligns with the targeted critical thinking skill indicator.	1 = Appropriate; 0 = Not appropriate

The maximum score for each question, if it meets all criteria, is 5. In addition to assigning a score, validators are asked to provide qualitative comments and suggestions for revision for each item deemed in need of improvement.

Data Analysis

Data analysis was conducted in two forms: descriptive analysis and quantitative analysis of content validity (Aiken, 1985; Shafie et al., 2022; Widyaningsih et al., 2021).

Descriptive analysis

The data from the validator assessment were analyzed descriptively to characterize the quality of each question item across the five assessed aspects. This analysis was used to identify general trends in the instrument's quality, determine which items performed well, and identify items that needed revision.

Content validity analysis with Aiken's V

The validity of each question item was calculated using Aiken's V (Aiken, 1985). Each question was assessed by five validators based on five assessment aspects: Ambiguity of Sentence Grammar (ASG), Ease of Understanding Question Intent (EUQ), Suitability of Breadth and Depth of Content (SBC), Accuracy of the Answer Key (AAK), and Conformity of the Question with the Critical Thinking Indicator (CQC). Each validator assigned a score of 1 if the aspect was appropriate (valid) and 0 if the aspect was inappropriate (invalid). The score for each question was calculated by summing the scores assigned to the five assessment aspects. For example, if Validator 1 assigned a score of 1 for all five aspects (ASG = 1, EUQ = 1, SBC = 1, AAK = 1, and CQC = 1), the total score for that question from Validator 1 was 5.

This process is carried out by five validators, each of whom assigns a score for the question based on these five aspects. After that, the scores of the five validators are averaged to obtain an s-value, the average score across all validators. The formula for calculating Aiken's V is presented in Equation (1) (Widyaningsih et al., 2021).

$$V = \frac{\Sigma s}{[n(c-1)]} \tag{1}$$

where V is the Aiken's V coefficient, s is the score assigned by each validator minus the lowest possible score ($s=r-l0$), r is the score assigned by the validator, l0 is the lowest possible score on the rating scale, c is the number of rating categories, and n is the number of validators.

Worked example. To illustrate the computation, consider item 2, which received total scores of 4, 4, 2, 5, and 3 from the five validators (each total is the sum of the five 0–1 aspect ratings, so the possible range is 0–5). Here the lowest possible category is $l0 = 0$, the number of categories is $c = 6$ (scores 0 through 5), and the number of validators is $n = 5$. For each validator, $s = r - l0 = r$, giving s-values of 4, 4, 2, 5, and 3 with $\Sigma s = 18$. Applying Aiken's $V = \Sigma s / [n (c - 1)] = 18 / [5 \times (6 - 1)] = 18 / 25 = 0.72$. Because $0.72 \geq 0.60$, item 2 is classified as valid, although its comparatively lower coefficient signals the minor conceptual revision noted by the validators.

Aiken's value V is in the range of 0 to 1. The closer it is to 1, the higher the item's content validity. The interpretation criteria used are as shown in Table 2.

Table 2. Interpretation of the validity coefficient

Coefficient	Interpretation
$0.6 \leq V \leq 1.0$	Valid
$V < 0.6$	Invalid

In addition to the index value, the final decision on each item also considers qualitative comments from the validator, ensuring that the instrument's revision is based not only on numerical scores but also on substantive considerations. To maintain the validity of the procedure, all validators receive an instrument grid, an operational definition of each critical-thinking indicator, and the same assessment guidelines prior to validation. This step is taken to minimize differences in interpretation among experts regarding the aspects being assessed. All validation data are stored in a protected digital document and is accessible only to researchers. The identities of validators are kept confidential and used solely for academic research. Commentary data and validator scores are used solely for revision and scientific reporting.

Methodological Limitations

This study focuses on content validation through expert judgment, so the results have not demonstrated the instrument's empirical validity among junior high school students. Therefore, further research is still needed to test the instrument's psychometric characteristics through field trials, including internal consistency, difficulty level, discriminative power, and construct validity based on actual student responses.

RESULTS AND DISCUSSION

Application of Ennis's Critical Thinking Skills Framework in Critical Thinking Instruments

This critical thinking skills test draws on Ennis's framework, emphasizing five key abilities: basic clarification, basic support, inference, advanced clarification, and strategies and tactics. The instrument is designed to assess these components specifically related to temperature and heat concepts and to align with junior high school physics objectives. Each ability is further divided into sub-aspects and indicators that gauge students' critical thinking skills, from question identification to evidence-based decision-making. Table 3 details these aspects, sub-aspects, and indicators, along with the corresponding question numbers that evaluate each.

Table 3. Description of Aspects, Sub-Aspects, and Indicators of Critical Thinking Skills According to the Ennis's Framework

Aspects of CTS	Sub-Aspects of CTS	Indicator	Question Number
Basic Clarification	Focusing Questions	-	-

Aspects of CTS	Sub-Aspects of CTS	Indicator	Question Number
	Analyzing Arguments	Evaluate the arguments in temperature and heat experiments	13-17
		Ask and answer clarifying questions	1-12, 18-23
Basic Support	Considering Credibility	-	-
	Observing Observations	Assess observation data related to temperature and heat	24-27
Inference	Making Reductions/Inductions	-	-
	Making and Considering the Value of Decisions	Evaluate the decisions based on observational evidence	28-32
Advanced Clarification	Defining Terms	Evaluate the definitions of the terms temperature and heat	33-37
Strategies and Tactics	Deciding on an Action	Evaluate action decisions related to temperature and heat	38-42

The resulting blueprint (Table 3) shows how each general critical thinking aspect and sub-aspect was rendered into specific, measurable indicators and mapped to particular question numbers. This mapping demonstrates that all five Ennis aspects are represented across the item set, with the largest concentration of items on basic clarification and progressively fewer items on the higher-order inference and strategies-and-tactics aspects. An example of the resulting CTS-THT item is presented in Figure 2.

Aspect: Provides basic clarification
Sub-Aspects: Asking and answering clarifying questions and challenging questions
Indicators: Evaluating opinions/
Views

Question:

Ramzi held objects A and B. He felt that object A felt colder than object B. Regarding this phenomenon, two junior high school students argued as follows:

Budi: The temperature of object A must be lower than the temperature B.
Rudi: The temperature of object A is not necessarily lower than the temperature of object B.

Whose opinion is acceptable in your judgment? Give the basis for your judgment!

Fig.2. Example CTS-THT

Results of Expert Validation and Instrument Improvement

The assessment of the test instrument focuses on its validity. The validity of the test instrument is established by considering input from experts. The test instrument was validated by 5 experts based on 5 identified aspects. Table 4 presents a summary of validation results from five validators who evaluated the critical thinking skills test instrument on temperature and heat materials.

Table 4. Recapitulation of Expert Validator Feedback Results

Question Number	Number of Suggestion	Problem	Follow-up
2	4	The concept of thermogun vs mercury thermometer accuracy; the key and reason are not yet solid.	Revise the concept of measuring water temperature with mercury thermometers and thermoguns; adjust the answer key and scientific reason.
3	2	Mixing of the terms heat, thermal energy, and internal energy.	Clearly distinguish the concepts of heat, thermal energy, and internal energy; double-check redaction, content, and keys.
17	2	The concept of solar radiation and energy changes in matter.	Revise of the concept of solar radiation: energy comes as electromagnetic waves; heat arises when radiation is absorbed by matter.
33	2	Inconsistencies in macroscopic definitions in the problem and microscopic in the keys.	Match the context of the definition to the question and the answer key; do not mix the macroscopic and microscopic levels.
38	2	The answer key related to cooking oil fire is unsafe and conceptually inappropriate.	Revise the answer key substantively: the reason water enlarges fire is not because it contains oxygen; the safe solution is to shut off the fire source or use CO_2.

Based on the expert validation results, several important findings were identified that required improvements to the instrument. Qualitative analysis showed that although most of the questions met the validity criteria, some still needed improvement, particularly in question wording, answer keys, and contextual relevance. In a study by Masek and Yamin (2012), revisions to question wording were found to be very important for avoiding ambiguity and enabling assessment instruments to measure student understanding more accurately.

We found some questions ambiguous and in need of correction to make them clearer and easier for students to understand. For example, question 6 requires revision to avoid confusion about calorie transfer, and questions 27 and 28 require revision to eliminate sentence ambiguity. Previous research by Megananda et al. (2021) also highlighted the importance of question clarity in reaction to ensure instruments can measure students' critical thinking skills. In addition, the analysis of question clarity showed that some questions, such as number 13, were less contextually aligned with conditions in Indonesia and needed improvement to be more relevant to students' daily lives. This aligns with the findings of Suyasa & Divayana (2022), who showed that questions relevant to the local context are easier for students to understand and more effective at measuring their critical thinking skills in everyday situations.

Regarding the content's breadth and depth, some questions need to be revised to be more in-depth and better aligned with the learning objectives. For example, question number 16 needs to more clearly link changes in the form of substances to global warming, and question number 17 needs to be corrected to accurately explain solar radiation as electromagnetic waves, not heat that turns into light. Research has resulted in errors in measuring students' critical thinking skills. In terms of the accuracy of the answer key, some questions require substantial revisions to the answer key, such as question number 10, which needs to emphasize that, in addition to water, fat also freezes at low temperatures, and question number 38, which is incorrect in explaining the effect of water on fire. This revision of the answer key is important to ensure the instrument accurately measures students' understanding, in line with Widyaningsih et al. (2021), who emphasize the need for accurate answer keys to effectively measure critical thinking skills.

Finally, regarding the suitability of the questions for the critical thinking indicator, some questions, such as question 4, can be improved by incorporating comparators' opinions to broaden the scope for evaluating arguments and to encourage students to think more critically. This is supported by Olivares-Rodríguez et al. (2017), who emphasize the importance of providing space to analyze and evaluate arguments so that students can think critically when facing questions. Based on the assessment results from five validators, the items in the critical thinking skills test instrument for the Temperature and Heat material can be grouped into three validity categories: valid without revision, valid with minor revision, and valid with substantive revision (Table 5). This grouping is based on a qualitative review of sentence structure, question clarity, content suitability, answer key accuracy, and alignment with critical thinking indicators. In general, the validation results indicate that most questions meet the expected validity criteria, so the developed instrument has a solid foundation for use after certain items are improved.

Table 5. Qualitative Validity Status of Question Items

Validity Status	Question Number	Remarks
Valid without revision	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 34, 35, 36, 37, 39, 40, 41, 42	The question is appropriate
Valid with minor revisions	17, 30, 33	Need to clarify concepts or improve answer keys
Valid with substantive revision	2, 38	Need to improve the concept and explanation of the answer

Of the 42 validated questions, 37 were deemed valid without revision. The questions in this category were appropriate in terms of redaction, clarity of intent, accuracy of concepts, answer key, and the representativeness of critical-thinking indicators. In some items, validators provided minor notes, but these were purely technical or editorial refinements and did not affect the substance of the questions. This aligns with the findings of Prasojo et al. (2020), who showed that most questions validated by experts can be used after minor revisions, as the revisions are limited to technical refinement and do not change the substance of the questions.

Furthermore, three questions are deemed valid with minor revisions: questions 17, 30, and 33. These questions meet the validity criteria but still require limited improvements to achieve optimal quality. The revisions mainly involve clarifying concepts or improving the answer key. For example, in question 17, the explanation of light energy still needs clarification to avoid misconceptions. Similar findings were discussed by Yahfizham et al. (2021), who emphasized the importance of clarifying concepts to help students understand the question. For question 30, the answer key needs revision to explain the effect of salt on ice. For question 33, the answer key needs improvement to align the definition of conduction with the concept, making it more conceptually accurate. This is reminiscent of Masek and Yamin (2012), who showed that questions with clearer and more structured explanations can help students develop their critical thinking skills.

In addition, two questions are valid with substantive revisions: questions 2 and 38. These two questions still have potential but require fundamental improvements in concepts and answer accuracy. Question 2 requires substantive revision because the explanation of how thermoguns are used to measure water temperature is scientifically incorrect; therefore, the answer key and supporting reasons need to be revised. This underscores that developing valid questions must prioritize scientific accuracy, which is essential for assessment instruments. This finding is also supported by Shafie et al. (2022), who emphasize the importance of conceptual correctness in assessment questions to ensure that the instrument accurately measures students' understanding. Meanwhile, question 38 requires a more thorough revision because its explanation of water's impact on fire remains conceptually incorrect and could confuse students. Susiyawati et al. (2021) also demonstrate that conceptual errors in questions can undermine the instrument's validity and lead to less precise assessments of students' critical thinking skills.

Overall, the qualitative analysis indicates that the developed test instruments have good validity, as most questions were deemed valid without revision. These findings are consistent with research by Suyasa et al. (2019), which showed that most assessment instruments attain adequate content validity after expert review, and only a few require minor or substantive revisions. This indicates that the developed instruments are aligned with the critical thinking skills indicators and the Temperature and Heat materials. However, some areas need improvement to enhance measurement accuracy, scientific rigor, and consistency. Additionally, the validity of each item in the critical thinking skills test instrument was evaluated using Aiken's V, as shown in Table 6.

Table 6. Validity based on Aiken's V Score

No.	V1	V2	V3	V4	V5	Aiken's V	No.	V1	V2	V3	V4	V5	Aiken's V
1	5	4	5	4	5	0.92	22	5	5	5	5	5	1.00
2	4	4	2	5	3	0.72	23	5	4	5	5	5	0.96
3	5	5	3	3	5	0.84	24	5	5	5	5	5	1.00
4	5	5	4	5	5	0.96	25	5	5	5	5	5	1.00
5	5	5	4	5	5	0.96	26	5	5	5	5	5	1.00
6	5	5	4	4	5	0.92	27	5	5	5	5	5	1.00
7	5	5	5	4	5	0.96	28	5	5	5	5	5	1.00
8	5	5	5	5	5	1.00	29	5	5	5	3	5	1.00
9	5	5	5	5	5	1.00	30	4	5	5	5	5	0.92
10	4	5	5	5	5	0.96	31	5	4	5	5	5	0.96
11	5	5	5	5	5	1.00	32	5	5	5	5	5	1.00
12	5	5	5	5	5	1.00	33	4	4	5	5	5	0.92
13	5	4	5	5	5	0.96	34	4	4	5	5	5	0.96
14	5	5	5	5	5	1.00	35	5	5	5	5	5	1.00
15	5	5	5	4	5	0.96	36	5	5	5	5	5	1.00
16	5	5	5	5	5	1.00	37	5	5	5	5	5	1.00
17	4	5	5	5	3	0.88	38	4	5	5	5	4	0.92
18	5	5	5	5	5	1.00	39	5	5	5	5	5	1.00
19	5	5	5	5	5	1.00	40	5	5	5	5	5	1.00
20	3	5	5	5	5	0.92	41	5	5	5	5	5	1.00
21	4	5	5	5	5	0.96	42	5	5	5	5	5	1.00

Based on Aiken's V calculations for 42 questions on a critical thinking skills test instrument related to Temperature and Heat, all questions were valid, with Aiken's V values close to or reaching 1.0. Aiken's V score was calculated from assessments by five validators who evaluated five main aspects, namely Sentence Ambiguity (ASG), Ease of Understanding Question Intent (EUQ), Suitability of Breadth and Depth of Content (SBC), Accuracy of Answer Key (AAK), and Conformity of the Question with Critical Thinking Indicators (CQC). Research by Wesnawa et al. (2022) also used the Gregory Formula to measure the content validity of digital instruments used to assess the pedagogical competence of prospective teachers, suggesting that using the right formulas, such as Aiken's V and the Gregory Formula, is critical to ensuring that the instruments used in education meet valid measurement standards.

Of the 42 questions evaluated, all had an Aiken's V ≥ 0.6, indicating they were valid. Research by Joseph et al. (2025) also found that high CVI values for teaching instruments, such as those tested on the Bvee kit, indicated that the instruments were highly valid and could be used directly without modification. This high validity indicated that most of the questions tested met the desired validity criteria, including editorial clarity, suitability with the material, and the ability to measure students' critical thinking skills. In addition, some questions that achieved a perfect Aiken's V score of 1.0, such as questions 8, 9, 11, 12, and 14, showed that these questions were very valid and could be retained without further content revision. The findings of Megananda et al. (2021) show that questions with an Aiken's V score of 1.0 can be retained without further content revision because they are fully aligned with the measurement objectives and relevant to the context being measured.

To complement the item-level indices, Figure 3 visualizes the distribution of Aiken's V coefficients across the 42 items and the pattern of validator ratings. The distribution confirms that item validity is strongly skewed toward the upper range, with 23 items achieving a perfect coefficient of 1.00 and only three items (2, 3, and 17) falling below 0.90. The heat map shows that the few lower scores are concentrated in a small number of validator-item cells rather than reflecting systematic disagreement.

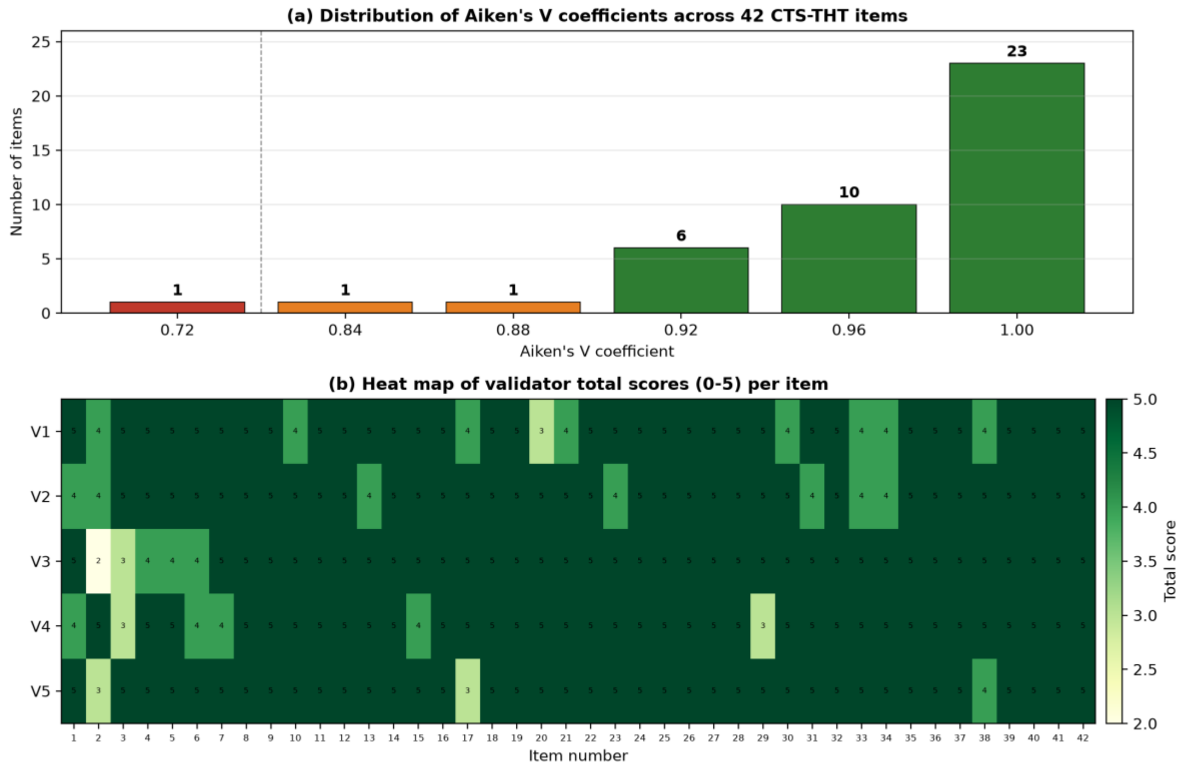

Fig.3. Distribution of Aiken's V coefficients and validator scores; a) Distribution of Aiken's V coefficients across 42 CTS-THT items, b) Heat map of validator total scores per item

However, two questions have Aiken's V scores between 0.6 and 0.8. Based on expert validators' suggestions, questions 2 and 21 require minor revisions to clarify concepts or refine wording to better align with students' measurement and comprehension goals. Prasojo et al. (2020) also suggest that questions with moderate validity scores should undergo minor technical improvements, especially to clarify concepts, so that instruments can better measure students' critical thinking skills.

Overall, the results of the analysis indicate that this test instrument has excellent validity, with all 42 questions demonstrating validity. These findings are consistent with research by Wesnawa et al. (2022), which affirms the importance of content validity in developing educational instruments to ensure they effectively measure the desired competencies. This study also reinforces the findings of Prasojo et al. (2020), who emphasized that technical refinements to questions with moderate validity do not change the substance of the questions but increase the clarity of the concepts measured.

Interpreted through Ennis's (1985) framework, the pattern of validator judgments is instructive. Items targeting basic clarification, such as asking and answering clarifying questions and analyzing arguments, were the most straightforward to validate because they require students to identify and restate key information in a physical situation, a relatively bounded cognitive demand that leaves little room for conceptual ambiguity. In contrast, items targeting inference and strategies-and-tactics were most frequently flagged for substantive revision (e.g., items 2 and 38) because they require chaining several reasoning steps: interpreting experimental evidence, generalizing about temperature and heat processes, and justifying a decision. Any imprecision in the scenario or answer key then propagates into the reasoning being assessed. This indicates that inference-oriented indicators are inherently more demanding to operationalize than clarification-oriented ones. For science learning, the implication is that instruction and assessment should explicitly scaffold inference and decision-making by making evidence, assumptions, and chains of reasoning visible, rather than assuming that students who can clarify a problem can also infer and decide correctly.

The consistently high Aiken's V values can be attributed to several procedural safeguards: all validators received the same instrument grid, operational definitions of each Ennis indicator, and identical scoring guidelines before rating, which minimized interpretive divergence; and every item was derived from an explicit blueprint that mapped it to a specific critical thinking indicator and a temperature-and-heat subtopic. Substantively, these results imply that the CTS-THT items are conceptually aligned with both the Ennis framework and the junior high school curriculum, allowing teachers to use the instrument to diagnose specific critical thinking indicators rather than a single global score; the main contribution to science education is therefore a curriculum-aligned, framework-based item bank whose content relevance has been established transparently at the item level. Nevertheless, an important limitation must be emphasized: content validity established through expert judgment attests only to the relevance and representativeness of the items and is not equivalent to construct validity. High Aiken's V values do not by themselves guarantee that the items function as intended when administered to students; evidence of construct validity, internal structure, and reliability must still be obtained through empirical field testing.

Situated within the broader science education and educational measurement literature, these results extend prior instrument-development work in two ways. First, they corroborate evidence that domain-specific critical thinking assessments, rather than general critical thinking batteries such as the Watson–Glaser Appraisal or the Halpern Assessment (Bernard et al., 2008; Butler, 2012), are needed to capture reasoning bound to particular science content, echoing Sartono et al. (2025) and Susiyawati et al. (2021). Second, the item-level Aiken's V evidence aligns with measurement scholarship emphasizing that content validity is necessary but not sufficient for valid inference (Shafie et al., 2022; Widyaningsih et al., 2021): expert agreement secures relevance and representativeness, yet the defensibility of test-score interpretations ultimately depends on subsequent empirical evidence. Interpreted this way, the CTS-THT contributes a theoretically grounded, content-validated foundation on which cognitive and psychometric claims can later be built, rather than a finished measurement claim.

Relevance of Instruments to Junior High School Physics Curriculum and Learning

This instrument is highly relevant to the junior high school science curriculum, especially for the topics of temperature and heat. In the junior high school science curriculum, developing critical thinking skills is a primary goal, particularly for understanding more complex scientific concepts such as heat transfer, temperature, and thermal energy. The importance of critical thinking development among junior high school students has also been emphasized in science education research (Mu'aziyah et al., 2023). This highlights the need for assessment instruments that can capture students' critical thinking abilities alongside their conceptual understanding. In the context of physics education, Sartono et al. (2025) noted that assessment instruments used in physics teaching, particularly for the topics of temperature and heat, should measure students' conceptual understanding in depth and encourage them to analyze and solve more complex physics problems.

The developed instrument measures students' critical thinking abilities in a more structured, in-depth manner. It aligns with the objectives for developing critical thinking skills expected in the junior high school curriculum. Susiyawati et al. (2021) revealed that, to measure students' critical thinking skills accurately, the instrument must be designed to identify and assess students' ability to think logically, especially in the context of difficult physics concepts, such as those related to temperature and heat. These instruments not only test students' conceptual understanding but also encourage students to analyze phenomena, assess arguments, and make decisions based on existing evidence. Research by Pina et al. (2021) supports this, showing that technology-based assessment instruments, such as those used in LMS, can encourage students to think critically and apply their knowledge to solve real-world problems in physics and engineering. This is especially important when teaching often challenging topics, such as temperature and heat, which require deep critical thinking skills to understand existing scientific concepts.

Enhancement Based on Validator Feedback

Most of the questions in this instrument have demonstrated high validity, but some require further improvement based on input from validators. For example, question 13 on the diversity of opinions about temperature and heat needs to be adjusted to better align with the Indonesian context, so students can more easily relate their physics knowledge to their daily experiences. Questions that receive substantial feedback, such as questions 2 and 38, require stronger physics concepts to ensure there are no misconceptions that could hinder the accurate measurement of students' critical thinking skills.

Limitations of Instruments and Advanced Research

Although this instrument has demonstrated high validity during expert validation, further research is still needed to test it empirically in a population of junior high school students. Field trials will allow further measurement of the instrument's reliability and its ability to assess students' critical thinking skills in real-world settings. Further research is also important to test the instrument's psychometric characteristics, such as question difficulty, discriminative power, and construct validity, in greater depth.

CONCLUSION AND SUGGESTION

This research successfully developed and validated a Critical Thinking Skills Assessment Instrument for junior high school students on the topic of heat and temperature, grounded in Ennis's (1985) critical thinking framework. The instrument was constructed to measure five core aspects of critical thinking: clarification, basic support, inference, advanced clarification, and strategies and tactics. Expert validation by five specialists indicated that the instrument has strong content validity: 37 items were judged valid without revision, while several items still require minor or substantial improvements, particularly in conceptual precision and wording clarity.

This study offers both theoretical and practical contributions. Theoretically, it provides an integrated framework that operationalizes Ennis's (1985) critical thinking abilities into physics-assessment indicators for temperature and heat, and it contributes item-level content-validity evidence clarifying which reasoning demands are easier or harder to operationalize. For science education research, the validated blueprint and item bank provide a reusable basis for developing and comparing domain-specific critical-thinking measures and for subsequent psychometric modeling. In practice, for classroom assessment, teachers can use content-validated items once their psychometric properties are confirmed to diagnose specific critical-thinking indicators rather than a single global score, thereby informing targeted instruction on temperature and heat. These contributions should nonetheless be read alongside the study's principal limitation: content validity is not equivalent to construct validity, and comprehensive psychometric analysis is required before the instrument is deployed for educational use.

Future studies are recommended to conduct systematic field testing of the instrument with junior high school students to examine its performance in authentic classroom contexts. Subsequent research should also investigate the instrument's reliability and apply Rasch modeling or other Item Response Theory (IRT) approaches to analyze item characteristics and scale functioning in greater depth. In addition, further work is needed to establish construct validity through empirical studies linking students' test scores to other indicators of critical thinking. It is also suggested that the instrument be adapted and extended to other physics topics for broader use in assessing critical thinking skills and ensuring stronger alignment with the overall science curriculum.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS STATEMENT

Conceptualization, S.R. and A.S.; methodology, S.R.; formal analysis, S.R.; theoretical framework development, A.S.; supervision, A.S., M.M., and I.K.; project administration, M.M.; resource management, I.K.; writing—original draft, S.R.; writing—review and editing, S.R. and M.Z.B.; critical review, commentary, and substantive input, M.Z.B.

ACKNOWLEDGMENTS

The authors sincerely thank the BPI Scholarship (Indonesian Education Scholarship), managed by the Center for Higher Education Funding and Assessment (PPAPT) and the Center for Education Financing Service (Puslapdik), with funding from the Indonesia Endowment Fund for Education (LPDP). This support was crucial for various research activities, including data collection, analysis, and manuscript writing. They also appreciate the academic environment and support that helped them complete this study successfully.

DECLARATION OF GENERATIVE AI SOURCES

During the preparation of this work, the author(s) used Scopus AI and Perplexity AI to assist with reference search and preliminary data analysis. After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Bernard, R. M., Zhang, D., Abrami, P. C., Sicoly, F., Borokhovski, E., & Surkes, M. A. (2008). Exploring the structure of the Watson-Glaser Critical Thinking Appraisal: One scale or many subscales? *Thinking Skills and Creativity, 3*(1), 15–22. https://doi.org/10.1016/j.tsc.2007.11.001

Butler, H. A. (2012). Halpern critical thinking assessment predicts real-world outcomes of critical thinking. *Applied Cognitive Psychology, 26*(5), 721–729. https://doi.org/10.1002/acp.2851

Dewi, S. E., Setyosari, P., & Hanafi, Y. (2025). Effect of inquiry-based learning on students' critical thinking: A meta-analysis. *Journal of Institutional Research South East Asia, 23*(3), 18–39.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*, 44–48.

Ennis, R. H. (1993). Critical thinking assessment. *Theory into Practice, 32*(3), 179–186. https://doi.org/10.1080/00405849309543594

Gavronskaya, Y., Larchenkova, L., Berestova, A., Latysheva, V., & Smirnov, S. (2022). The development of critical thinking skills in mobile learning: Fact-checking and getting rid of cognitive distortions. *International Journal of Cognitive Research in Science, Engineering and Education, 10*(2), 51–68. https://doi.org/10.23947/2334-8496-2022-10-2-51-68

Ghanizadeh, A., Bajestani, G. S., Hosseinpour, F., Hosseini, A., Makhloughi, F., & Toosi, M. B. (2024). Mindfulness-enhancing instruction (MEI): Contributions to electroencephalogram (EEG) dynamics, higher order thinking skills (HOTS), and effective learning. *Thinking Skills and Creativity, 53*, 101611. https://doi.org/10.1016/j.tsc.2024.101611

Illene, S., Feranie, S., & Siahaan, P. (2023). Create multiple-choice tests based on experimental activities to assess students' 21st century skills in the heat and heat transfer topic. *Journal of Education and Learning (EduLearn), 17*(1), 44–57. https://doi.org/10.11591/edulearn.v17i1.20540

Joseph, V., Shaharudin, S. M., Hamid, N. A., & Rajak, N. A. (2025). Development and evaluation of the BVEE kit for teaching circle topics: A validity and usability study in Form 2 mathematics. *International Journal of Information and Education Technology, 15*(2). https://doi.org/10.18178/ijiet.2025.15.2.2236

Masek, A., & Yamin, M. (2012). The impact of instructional methods on critical thinking: A comparison of problem-based learning and conventional approach in engineering education. *International Scholarly Research Network, 2012*, Article 759241. https://doi.org/10.1155/2012/759241

Megananda, A., Muzayyanah, E., Darmayanti, H. P., & Priana, Z. I. (2021). Development of digital distance measurement instrument based on Arduino Uno for physics practicum. *IMPULSE: Journal of Research and Innovation in Physics Education, 1*(2), 80–88. https://doi.org/10.29303/impulse.v1i2.85

Mu'aziyah, S. E. S., Hidayat, T., Sriyati, S., & Lutianasari, L. (2023). Implementation of the Merdeka Curriculum using Citizen Science Project Weather-it to improve critical thinking skills of junior high school students. *Jurnal Penelitian Pendidikan IPA, 9*(3), 1470–1479. https://doi.org/10.29303/jppipa.v9i3.2277

Nurhayati, N., Wahyudi, W., Suhandi, A., Muslim, M., Kaniawati, I., & Saputri, D. F. (2026). Enhancing prospective teachers' skills in designing physics learning: Developing electronic learning materials with interactive reading, scaffolding and modeling (IRTaMS) strategies. *IJIS Edu: Indonesian Journal of Integrated Science Education, 8*(1), 27–45.

Olivares-Rodríguez, C., Guenaga, M., & Garaizar, P. (2017). Automatic assessment of creativity in heuristic problem-solving based on query diversity. *Dyna, 92*(4), 449–455. https://doi.org/10.6036/8243

Parkkinen, V.-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M. P., Norell, C., Russo, F., Shaw, B., & Williamson, J. (2018). Evaluating evidence of mechanisms. In *Evaluating evidence of mechanisms in medicine* (pp. 77–90). Springer. https://doi.org/10.1007/978-3-319-94610-8_6

Pina, J., Ellis, G. W., Rudnitsky, A., Mazur, R., McGinnis-Cavanaugh, B., & Huff, I. (2021, July). *Developing a measure to capture middle school students' interpretive understanding of engineering design*. ASEE Virtual Annual Conference Content Access. https://doi.org/10.18260/1-2--36934

Popescu, A., & Morgan, M. (2007). Teaching information evaluation and critical thinking skills in physics classes. *The Physics Teacher, 45*(8), 507–510. https://doi.org/10.1119/1.2798365

Pramasdyahsari, A. S., Setyawati, R. D., Aini, S. N., Nusuki, U., Arum, J. P., Astutik, L. D., Widodo, W., Zuliah, N., & Salmah, U. (2023). Fostering students' mathematical critical thinking skills on number patterns through digital book STEM PjBL. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(7). https://doi.org/10.29333/ejmste/13342

Prasojo, L. D., Wijayanti, W., Yuliana, L., Agus, N., Habibi, A., & Yaakob, M. F. M. (2020). Instruments' validation of access to motivation, skills, and use of digital technology: EFL context in Indonesia. *Studies in English Language and Education, 7*(2), 308–322. https://doi.org/10.24815/siele.v7i2.16788

Rahayu, S., & Alsulami, N. M. (2024). Assessing higher order thinking skills of the 21st century learners using socio-scientific issues as a context. *AIP Conference Proceedings, 3106*(1), 070009. https://doi.org/10.1063/5.0214793

Saputri, M., Saminan, S., & Nisa, Z. (2026). Differences in students' critical thinking and collaboration skills between problem-based learning and discovery learning in physics. *Equator Science Journal*, *4*(2), 210–222. https://doi.org/10.61142/esj.v4i2.372

Sartono, B., Sunarno, W., Prayitno, B. A., & Indriyanti, N. Y. (2025). Development and validation of a critical thinking assessment on temperature and heat for secondary physics education. *Educational Process: International Journal, 16*, 264. https://doi.org/10.22521/edupij.2025.16.264

Seibert, S. A. (2021). Problem-based learning: A strategy to foster Generation Z's critical thinking and perseverance. *Teaching and Learning in Nursing, 16*(1), 85–88. https://doi.org/10.1016/j.teln.2020.09.002

Shafie, H., Majid, F. A., & Ismail, I. S. (2022). Developing a 21st century technological pedagogical content knowledge (TPACK) instrument: Content validity and reliability. *International Journal of Education, 14*(3), 100–115. https://doi.org/10.5296/ije.v14i3.19980

Susiyawati, E., Sudibyo, E., & Sari, D. A. P. (2021). Development and validation of an instrument for assessing middle school students' critical thinking skills. *International Journal of Assessment and Evaluation, 28*(2), 1–13. https://doi.org/10.18848/2327-7920/CGP/V28I02/1-13

Suyasa, P. W. A., & Divayana, D. G. H. (2022). Instrumen evaluasi model discrepancy-CSE-UCLA dalam rangka menunjang evaluasi efektivitas pelaksanaan pembelajaran synchronous. *Jurnal Pedagogi dan Pembelajaran, 5*(2), 197–207. https://doi.org/10.23887/jp2.v5i2.48447

Suyasa, P. W. A., Kurniawan, P. S., Ariawan, I. P. W., Sugandini, W., Adnyawati, N. D. M. S., Budhyani, I. D. A. M., & Divayana, D. G. H. (2019). Empowerment of CSE-UCLA model based on Glickman quadrant aided by visual application to evaluate the blended learning program on SMA Negeri 1 Ubud. *Journal of Theoretical and Applied Information Technology, 96*(18), 6203–6219. https://doi.org/10.33193/JATIT.2020.96.18.6203

Velmovská, K., & Bartošovič, L. (2016). Developing critical thinking skills in physics classes. In *Critical thinking: Theories, methods and challenges* (pp. 1–43).

Wei, J. H., Chuang, H. H., & Smith, T. J. (2022). The relationship between a school culture's openness to creative solutions and inquiry-based teaching practices. *Journal of Creative Behavior, 56*(3), 382–395. https://doi.org/10.1002/jocb.535

Wesnawa, I. G. A., Kartowagiran, B., Jaedun, A., Hamdi, S., Hadi, S., Sunendar, D., Abdul, L., Laliyo, R., Christiawan, P. I., & Gede, D. (2022). Content validation of digital instrument for measurement of pedagogic competence for social science teacher candidates in the Industrial Revolution 4.0 era in Indonesia. *International Journal of Information and Education Technology, 12*(12), 1424–1430. https://doi.org/10.18178/ijiet.2022.12.12.1767

Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2021). The development of the HOTS test of physics based on modern test theory: Question modeling through e-learning of Moodle LMS. *International Journal of Instruction, 14*(4), 51–68. https://doi.org/10.29333/iji.2021.1444a

Yahfizham, I., Yusti, I., & Hamzah, M. L. (2021). The test of construct validity for one-factor model. *Journal of Educational Research and Evaluation, 5*(3), 429–435. https://doi.org/10.15294/jered.v5i3.4093