

Research Article

Cite this article: 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, e2025264.
<https://doi.org/10.22521/edupij.2025.16.264>

Received April 17, 2025

Accepted May 25, 2025

Published Online June 20, 2025

Keywords: Critical thinking test, temperature and heat, instrument development, physics education

Author for correspondence:

Widha Sunarno

✉ widhasunarno@staff.uns.ac.id

✉ Universitas Sebelas Maret, Indonesia.



OPEN ACCESS

© The Author(s), 2025. This is an Open Access article, distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction, provided the original article is properly cited.

Development and Validation of A Critical Thinking Assessment on Temperature and Heat for Secondary Physics Education

Bangun Sartono, Widha Sunarno, Baskoro Adi Prayitno, Nurma Yunita Indriyanti

Abstract

Background/purpose. Critical thinking (CT) is fundamental in science education, but instruments to measure CT in specific domains, such as physics, are still limited. The present study aims to develop and validate the Critical Thinking Test in Temperature and Heat (CTTH), an instrument designed to measure critical thinking skills in the topic of temperature and heat in physics.

Methods. The development of CTTH refers to a common critical thinking test framework. The validation process involved expert review from seven experts in physics education, a small-scale pilot test with 33 students, and final validation with 720 secondary school students, ensuring the test items' clarity, relevance, and psychometric quality. Instrument reliability was measured using Cronbach's alpha for internal consistency and Fleiss' kappa for inter-rater reliability.

Results. The CTTH exhibits adequate reliability, with internal consistency and inter-rater reliability, confirming the instrument's effectiveness in measuring critical thinking skills in the context of physics. CTTH is a measurement instrument that can measure critical thinking skills related to temperature and heat in physics learning.

Conclusion. The CTTH instrument offers a resource for further research on the integration of critical thinking in physics education. Future research is suggested to expand the critical thinking assessment framework and answer research questions for wider application in science education.

1. Introduction

Critical thinking (CT) is a central goal in science education, enabling students to think rationally, analytically, and reflectively when faced with scientific problems. Numerous studies have demonstrated that integrating CT into science instruction improves students' understanding of scientific concepts and supports informed decision-making in real-life situations (Affandy et al., 2024; Arnold & Millar, 1994; Bajracharya et al., 2019). Responding to these demands, educational reforms increasingly emphasize CT development, particularly in physics education, where students often struggle to connect abstract concepts with practical applications (Arnold & Millar, 1994; Cari et al., 2020; Nurhuda et al., 2017).

Although the importance of CT is widely acknowledged, assessing it remains challenging. Disagreement about its definition has resulted in diverse assessment approaches with varying emphasis and effectiveness (Alpizar et al., 2022; Hsu et al., 2022). Furthermore, most existing instruments—such as the Watson-Glaser and Cornell CT tests—focus on domain-general skills, making them less sensitive in measuring how students apply CT in discipline-specific contexts like physics (Casella et al., 2020; Engelhard & Wang, 2021; Zehirlioglu & Mert, 2020). This issue highlights a significant gap: while general CT skills are essential, they may not fully capture the nuanced reasoning required in specific fields (Häkkinen et al., 2025).

The distinction between domain-general and domain-specific approaches to CT has become increasingly relevant in science education (Kassiavera et al., 2024; Mafinejad et al., 2017). Domain-specific assessments, tailored to the content and cognitive demands of a subject like physics, are believed to provide more valid insights into students' actual thinking processes (Kassiavera et al., 2024; Mafinejad et al., 2017). However, few instruments exist that measure CT within specific physics topics, such as temperature and heat—despite these being foundational yet conceptually challenging areas in early college-level physics (Kassiavera et al., 2024; Suwita et al., 2024).

To address this gap, the present study proposes the development and validation of a Critical Thinking Test in Temperature and Heat (CTTH). Unlike existing general CT instruments, the CTTH is designed to assess both general and domain-specific critical thinking skills, based on Facione (2000) six components: interpretation, analysis, evaluation, inference, explanation, and self-regulation. This study aims to contribute to the field by offering a valid and reliable instrument for measuring CT in the context of physics—specifically on the topic of temperature and heat—thereby enhancing efforts to promote and assess CT in science education.

Accordingly, this study addresses the following research questions: (1) What is the validity and reliability of the CTTH instrument in measuring students' critical thinking skills on the topic of temperature and heat in physics?, (2) How well do CTTH items discriminate between students with high and low critical thinking skills?, and (3) To what extent can the CTTH measure each of the six components of critical thinking in the context of temperature and heat?.

2. Literature Review

2.1. Critical Thinking in Physics Education

A review of five commonly used domain-general CT tests (as presented in Table 1) was conducted to evaluate whether the definitions given in each test were clear, whether the skills measured were general and applicable across disciplines, and whether the items in the tests were valid in measuring the intended critical thinking skills. The literature review aimed to identify elements of previously developed tests that could be adapted or refined to construct a critical thinking assessment specifically focused on temperature and heat in physics.

Table 1. Description of commonly administered domain-general CT tests

CT instrument		CT components	Item format	Reference
Cornell Thinking Test Level X/Z	Critical	Induction, deduction, credibility, identification of assumptions	Multiple choice	(Leach et al., 2020)
Facione's Thinking Disposition and Skills Instrument	Critical	Interpretation, analysis, evaluation, inference, explanation, and self-regulation	Multiple choice	(Facione, 2000)
Ennis-Weir Thinking Essay Test	Critical	Argument analysis, recognizing fallacies, making judgments	Essay (open-ended)	(Ennis, 1958)
Watson-Glaser Thinking Appraisal	Critical	Inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments	Multiple choice	(Bernard et al., 2008)
Lawson's Thinking in Science Test (CTSR)	Critical	Scientific reasoning (proportional thinking, probabilistic thinking, correlational thinking, hypothetico-deductive reasoning)	Multiple choice	(Lawson, 1992)

The review of various CTs showed that Facione's Critical Thinking Disposition and Skills Instrument was chosen as the main reference in developing CT tests for temperature and heat. The test was chosen because it has several advantages, including the relevance of the CT measured in the variety of question formats used and the broad representation of various real-life contexts. The skills measured in this instrument, such as interpretation, analysis, evaluation, inference, explanation, and self-regulation, are especially relevant in physics education, particularly in helping students develop CT that can be applied in everyday situations and in solving more complex problems.

Table 2. Elaboration of the domain-specific CT outcomes targeted in the CTTT test

CT components	Domain-specific CT outcomes
In the context of temperature and heat, the student will be able to:	
Interpretation	Interpret temperature and heat data from physics experiments and relate it to relevant concepts.
Analysis	Analyze experimental temperature and heat change data and identify patterns or relationships between physics variables.
Evaluation	Evaluate the results of experiments on temperature and heat and assess the suitability between theory and experimental results.
Inference	Conclude the heat transfer phenomenon based on observations of temperature and energy changes in physical systems.
Explanation	Explain the heat transfer process in various contexts (conduction, convection, radiation) using appropriate terminology and physical principles.
Self-regulation	Correct their understanding and adjust experimental procedures to understand temperature and heat better.

The next phase in developing a CT test in the context of physics is to analyze the objectives of the items in Facione's Critical Thinking Disposition and Skills Instrument (Facione, 2000), with modifications using a two-tier test. Contextual adjustments are made by compiling learning outcome objectives more specific to the context of temperature and heat in high school students, as presented in Table 2. Adjustments are made so that the test does not cover all the temperature and heat topics content but only focuses on essential concepts relevant to developing students' CT. The main purpose of using this instrument is to measure students' CT in a domain-specific context after they have learned about temperature and heat. The test is not intended to be a comprehensive assessment of the entire content but rather to assess how students can apply the CT taught in relevant situations.

2.2. Two-Tier Test as a Instrument to Assess Critical Thinking

The two-tier test is one form of assessment instrument designed to measure students' conceptual understanding in more depth. According to Treagust (1988), a two-tier test consists of two parts for each item. The first tier is in the form of multiple-choice questions that ask for answers to a scientific problem or phenomenon. Meanwhile, the second tier asks for reasons or justifications for the answer chosen in the first tier, also in the form of multiple choice (Kassiavera et al., 2024; Suwita et al., 2024). Two-tier tests allow educators to not only know students' final answers but also understand the rationale or scientific reasoning behind their choices (Lawson, 1992; Sapia et al., 2022). Therefore, a two-tier test can reveal any misconceptions or weaknesses in students' reasoning, which may not be apparent if only a single multiple-choice item is used (Alwan, 2011; Zakwandi et al., 2024). The two-tier test format is widely used in science education because it integrates conceptual knowledge with CT, including logical reasoning and argument evaluation.

The two-tier format is suitable for uncovering students' CT and conceptual errors because it is able to evaluate not only the final answer but also the reasoning or thought process underlying students' choices. Therefore, in the context of critical thinking assessment, the important aspects that need to be assessed are students' ability to provide logical justifications, evaluate arguments, and draw appropriate inferences (Alwan, 2011; Lawson, 1992; Sapia et al., 2022). The two-tier structure provides space for students to reveal conceptual understanding as well as reasoning ability through the reasons chosen in the second tier. If a student gives the correct answer in the first tier but the reasoning chosen in the second tier is incorrect, then this may indicate a conceptual error or guesswork (Kaltakci et al., 2016). However, if the correct reason is chosen even though the final answer is wrong, then it reflects the potential for critical thinking that has not been properly facilitated (Cetin-Dindar & Geban, 2011). Accordingly, the two-tier format provides more comprehensive and diagnostic information on student thinking than conventional multiple-choice questions, making it highly relevant for use in the measurement of critical thinking skills.

2.3. Instrument Validation Using the Rasch Model

The Rasch Model contributes to analyzing the validity and reliability of a test instrument, especially in the context of measuring cognitive abilities such as critical thinking. The Rasch Model is an approach in item response theory (IRT) that allows an assessment of the extent to which each item in a test functions consistently and fairly across all respondents (Affandy et al., 2021; Kaur et al., 2024). The Rasch Model helps identify whether each item is able to measure the same construct unidimensionally and whether the difficulty level of the item matches the ability range of the participants (Wind, 2019). The Rasch Model also detects misfitting items, which can interfere with the interpretation of measurement results (Bond et al., 2021). Meanwhile, in terms of reliability, the Rasch model provides accurate estimates of measurement consistency at both the individual and item levels, such as through the person and item reliability indices (Cascella et al., 2020; Zehirlioglu & Mert, 2020). Rasch Model thus not only provides quantitative information about the characteristics

of items and participants but also ensures that the instruments used have accuracy and reliability in measuring CT in an objective and standardized manner.

The Rasch Model evaluates the fit between items and student ability through probabilistic analysis that compares actual response patterns with patterns predicted by the model. Each item is assumed to have a certain level of difficulty, while each student has a certain level of ability (Danday & Monterola, 2019; Mundilarto & Ismoyo, 2017). The model then calculates the probability of a student with a certain ability to answer an item with a certain difficulty level correctly. The fit between the empirical data and the model is assessed using statistics such as infit and outfit mean square, which indicate the extent to which students' responses to an item match the model's expectations (Karaca-Atik et al., 2023; Stolk et al., 2021). A mean square value close to 1 indicates that the item is functioning optimally. At the same time, values that are too high or too low may indicate that the item is unsuitable, for example, because it is too predictable or too ambiguous (Kinoshita, 2022; Suwita et al., 2024). Moreover, the Rasch Model also allows visualization of the match between student ability and item difficulty through the item-person map, which helps to ensure that the available items represent the entire spectrum of student ability.

The Rasch Model has a number of advantages over classical approaches in assessment instrument development, especially in terms of accuracy, objectivity, and generalizability of measurement results. However, unlike the CTT, which is highly dependent on the characteristics of the test taker group, the Rasch Model produces item parameter estimates and student abilities that are independent of each other (Cascella et al., 2020). It allows the measurement results to be more stable even when used in different populations (Cvenic et al., 2022). Moreover, the Rasch Model provides more detailed and diagnostic information, such as the level of item fit, item and respondent reliability, and item-person map of the distribution of student ability and item difficulty (Collado et al., 2015). Another advantage is its ability to identify problematic items statistically and logically, which is very helpful in the instrument revision and improvement process. The Rasch Model also supports the development of fair and equitable instruments, as it allows differential item functioning (DIF) analysis to assess whether an item shows bias towards a particular group (Ridho, 2018). Therefore, the Rasch Model is a comprehensive analytical tool for ensuring the validity and reliability of assessment instruments, especially in measuring CT.

3. Methodology

The present study aimed to develop and validate a CTT instrument for senior high school students. The initial stages included defining the critical thinking construct concerning Facione (2000) model, which includes six sub-skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation. The instrument was developed in a two-tier test format, which simultaneously allows the measurement of conceptual understanding and critical reasoning. A total of 30 items were developed based on expert review. The scoring guide was systematically organized using a four-category polytomous scale, which considered a combination of the accuracy of multiple-choice answers and the reasons given by students. Content validity was tested through assessment by seven physics experts, focusing on content appropriateness, critical thinking skills measured, and visual and verbal clarity of the items. A small-scale pilot test was conducted with 33 students to assess the items' clarity, the scoring guide's suitability, and the processing time. After revision, the main trial was conducted with 720 students using standardized administration procedures. Inter-rater reliability was analyzed using Fleiss' Kappa based on ratings from seven independent raters, while internal consistency was analyzed to ensure the instrument's reliability. Data were analyzed using the Rasch Model to evaluate item difficulty and discrimination and fit the theoretical model through infit and outfit mean square (MNSQ) analysis.

3.1. Item Format

Many existing CT tests use multiple-choice to assess students' critical thinking skills. Although its advantages include time efficiency and ease of administration, many critics claim that multiple choice is not effective enough to measure critical thinking skills in depth. Accordingly, multiple-choice tests factual knowledge or basic understanding without requiring students to actively analyze, evaluate, or apply learned concepts in a more complex context. In response to these limitations, it was decided that a two-tier test format would be used to develop CTTH. The two-tier test format allows for a more holistic assessment of students' CT, where the first item tests basic understanding or concepts, while the second item tests students' ability to apply that knowledge in a more complex context, thus providing a more comprehensive overview of students' CT.

3.2. Item Construction

The item construction process in the development of CTTH was gradual and iterative to ensure the suitability of the content to the domain context and high content validity. The process began with the initial development of 25 items based on the CT framework and the topic of temperature and heat. The items then underwent an internal review process by a team of experts by considering three criteria: compatibility with CT measurement objectives, clarity of question statements, and compatibility with the characteristics of high school students as the target population. Based on the review results, five new items were added by referring to concepts in relevant physics textbooks, and revisions were made to previous items to enhance their clarity and scientific context. Following this process, the final test format consisted of 30 items in the form of a two-tier test, with each item consisting of two levels of questions: the first level measuring concept understanding, and the second level measuring reasoning or justification that represents CT in the topic of temperature and heat.

3.3. Creating Scoring Guide

The development of the scoring guide in CTTH is based on the need to provide an objective, transparent, and accountable assessment of student performance in answering question items. The process of creating a scoring guide is carried out systematically through several stages, starting from the formulation of ideal answers to each item, identification of response features that reflect the quality of students' CT, to the determination of scoring weights based on the level of complexity and the estimated duration required by students to answer. The guide is designed to accommodate the variety of responses that test takers may give, so it does not only assess the correctness or incorrectness of the final answer but also pays attention to the quality of the underlying reasoning. The scoring system adopted a four-category polytomous scaling adapted from the research of Affandy et al., (2021). A score of 1 is given if the student is wrong in answering the multiple choice and also gives an incorrect reason, which reflects inaccuracy in the use of concepts and results. A score of 2 is given if the student chooses the correct multiple-choice answer, but the reason given is incorrect - this condition indicates that the student can obtain the correct result but not based on the right concept. Conversely, a score of 3 is given if the student chooses the wrong multiple-choice answer but gives the correct reason, which indicates that although the understanding of the concept is quite good, the application is not yet accurate. Finally, a score of 4 is given if the student can answer the multiple-choice correctly and give the correct reasoning, indicating a complete understanding and application of the correct concept.

3.4. Expert Review

Content validity and inter-rater reliability in developing the CTTH instrument were obtained through expert assessments conducted independently from the item construction process. Seven professors with expertise in physics education and critical thinking were invited as assessors. The selection criteria for these experts included: (1) holding a doctoral degree in physics or physics

education, (2) having a minimum of five years of experience in teaching or research related to science education and/or assessment, and (3) having published scholarly work on physics education, critical thinking, or assessment design. These experts evaluated the thirty developed items based on three aspects: the alignment between item objectives and the components of critical thinking skills, the accuracy of the content in terms of physics concepts, and the visual and verbal clarity of both question presentation and answer choices.

3.5. Small-Scale Paper-Pencil Administration

Initial empirical validation of the CTTH instrument was carried out through a small-scale field trial involving 33 students as research subjects. It aimed to evaluate participants' responses to each item, test the suitability of the scoring guide with the answers given, and assess the feasibility of the overall test processing time. The pilot test results showed that most of the items were well understood by the participants, but there were some unclear or ambiguous responses, both in the first tier (answer choices) and the second tier (reasons). The pilot test findings prompted revisions to the item formulation and adjustments to the scoring guide to make it more explicit and appropriate to the variations in participants' actual responses. The time required to complete the entire test ranged from 50 to 60 minutes, which was considered adequate for the level of complexity of the items presented.

3.6. Item Revision and Administration

The finalization phase of the CTTH instrument was conducted after an in-depth analysis of the initial pilot test results, focusing on revising items that showed ambiguity or weakness in measuring critical thinking (CT). The main trial involved 720 secondary school students selected through stratified random sampling to ensure representation across various geographical locations (urban, suburban, rural, and border between rural and urban), school types, and student backgrounds. This sampling strategy was intended to enhance the generalizability of the findings. Prior to participation, informed consent procedures were implemented. Permission was obtained from school administrators, and parental consent was secured for all student participants. Students were informed about the voluntary nature of participation, the purpose of the study, and the confidentiality of their responses. The administration of the instrument was conducted under close supervision, with standardized verbal instructions provided to all participants to ensure consistency. The estimated administration time was approximately one hour, adjusted according to the complexity and number of test items.

Table 3 presents the demographic characteristics of the respondents. According to gender, most participants were female (68.3%, 492 students), while males comprised 31.7% (228 students). Regarding age, most were 17 years old (72.36%), followed by 16-year-olds (26.67%), and a small proportion of 18-year-olds (0.97%). In terms of school location, participants were evenly distributed across cities (40%), suburbs (20%), villages (20%), and areas on the border of villages and cities (20%). This distribution reflects a variety of educational settings that may influence students' critical thinking abilities.

Table 3. Demographics of Respondents in the Study

Category	Sub category	Frequency	%
Gender	Male	228	31.7
	Female	492	68.3
Age	16 Years	192	26.67
	17 Years	521	72.36
	18 Years	7	.97
School affiliation	High School A	144	20
	High School B	144	20
	High School C	144	20
	High School D	144	20
	High School E	144	20
Geographic location	Urban	288	40
	Suburban	144	20
	Rural	144	20
	Border between rural and urban	144	20

3.7. Inter-rater reliability and Internal Consistency

Inter-rater reliability measures the level of consistency of scores given by more than one rater to students' open-ended responses, especially in the reasoning section in the two-tier format. Seven independent raters trained using the scoring guide scored the 30 CTH items separately. In order to measure inter-rater agreement, Fleiss' Kappa was used which is appropriate for situations with more than two raters (Fleiss, 1971). Meanwhile, internal consistency measures the extent to which the items in the instrument provide consistent results in measuring the construct of CT.

3.8. Item Difficulty and Discrimination

Item difficulty measures the difficulty of an item, i.e. the extent to which the item can be answered correctly by test takers. According to the Rasch Model, this parameter is measured by a log-odds value that describes the likelihood that a participant of average ability will be able to answer the item correctly (Cascella et al., 2020; Zehirlioglu & Mert, 2020). The optimal item difficulty value is in the range of challenging enough for most participants without being too easy or difficult (Bond et al., 2021). On the other hand, item discrimination measures an item's ability to distinguish between participants with different ability levels. Item discrimination in the Rasch Model is interpreted through the discrimination parameter coefficient, which reflects how well the item discriminates between high and low-ability participants. The higher the discrimination value, the better the item differentiates between participants with different CT levels.

3.9. Rasch Model Fit Analysis

Rasch model fit analysis is a procedure used to evaluate the extent to which the data obtained from the test fits the Rasch model used to measure test takers' ability. The analysis aims to ensure that the items and test taker responses conform to the basic assumptions of the Rasch Model, which

include unidimensionality, item independence, and accuracy of ability level measurement (Affandy et al., 2021; Kaur et al., 2024). The Rasch Model was chosen over other psychometric approaches because it provides item-level diagnostics, supports objective measurement through sample-independent item calibration, and is well-suited for validating educational assessments that aim to measure latent traits like critical thinking. Additionally, the Rasch approach allows the transformation of ordinal raw scores into interval-level data, enhancing the precision and interpretability of test scores (Bond et al., 2021).

Rasch Model fit analysis was conducted to evaluate the fit between the data and the theoretical model espoused by the instrument. The results of this analysis involved examining the infit mean square (MNSQ) and outfit MNSQ values for each item, which indicate the extent to which the item functions according to the model's expectations. MNSQ values within the range of 0.5 to 1.5 indicate a good fit (Wind, 2019), with values smaller than 0.5 indicating items that are too predictable and values greater than 1.5 indicating items that are too unpredictable or misfitting (Bond et al., 2021).

4. Results

4.1. Content Validity

Aiken's V index is used to evaluate the degree of agreement between raters on each item's relevance to the measured construct. Aiken's V values range from 0 to 1, with values close to 1 indicating a high level of content validity (Aiken, 1980). Based on the assessment results of 30 items in CTTH, Aiken's V index varies from .81 to .96. The highest value is found in item 25 ($V = .96$), while the lowest is in item 8 ($V = .81$). The overall average Aiken's V value was 0.88, indicating that each item generally had good content validity according to the experts' assessment. The values indicate that the raters have a high consensus in assessing that the items developed have adequately represented the critical thinking indicators. It provides the basis that the CTTH instrument has relevant content coverage and is based on Facione (2020) theoretical framework.

Table 4. Summary statistics for the CTTH items

Item	Content Validity	Inter-Rater Reliability	Item difficulty	Standard Error	Infit MNSQ (logit)	Outfit MNSQ (logit)	Item Discrimination
Item 1	.86	.85	.56	.13	1.01	1.00	.33
Item 2	.88	.86	-.36	.09	1.01	1.01	.37
Item 3	.88	.87	.40	.12	1.01	1.01	.42
Item 4	.91	.88	-.79	.09	.99	.98	.43
Item 5	.87	.86	-.68	.09	1.00	.99	.51
Item 6	.85	.89	-.51	.09	1.00	1.00	.60
Item 7	.82	.79	-.75	.09	.94	.89	.33
Item 8	.81	.80	1.60	.09	.98	.99	.64
Item 9	.86	.84	-.35	.10	1.03	1.01	.51
Item 10	.88	.81	-.42	.09	1.05	1.06	.40

Item 11	.89	.87	1.84	.11	.98	.93	.37
Item 12	.91	.88	-.56	.09	.98	.93	.25
Item 13	.87	.86	.48	.13	1.02	1.02	.49
Item 14	.85	.85	-.41	.09	1.11	1.01	.44
Item 15	.92	.86	-.78	.09	.94	.93	.42
Item 16	.83	.87	-.51	.09	1.08	1.02	.34
Item 17	.86	.88	-.82	.09	.92	.89	.36
Item 18	.88	.86	.39	.12	.99	.95	.38
Item 19	.89	.89	-.47	.09	1.01	1.01	.43
Item 20	.95	.79	-.48	.09	1.00	.93	.23
Item 21	.87	.80	-.51	.09	1.08	1.01	.32
Item 22	.85	.84	-.17	.10	1.05	1.07	.34
Item 23	.92	.81	-.69	.09	1.05	1.06	.35
Item 24	.91	.80	-.39	.09	.94	.88	.37
Item 25	.96	.84	.33	.12	.92	.71	.37
Item 26	.88	.81	1.15	.26	.96	.48	.55
Item 27	.89	.86	1.75	.22	1.03	1.03	.45
Item 28	.91	.85	.84	.15	.92	.62	.36
Item 29	.87	.88	1.07	.16	1.00	.90	.44
Item 30	.85	.86	1.47	.19	1.01	1.03	.46

4.2. Inter-Rater Reliability

Inter-rater reliability is an important indicator in ensuring that the instrument developed has consistent assessment from experts, especially in assessing the suitability of items with CT constructs. Inter-rater reliability is calculated using the Fleiss' Kappa statistic, which measures the degree of agreement of three raters on the relevance of each item to the six subscales of critical thinking, according to Facione (2020). Based on Table 4, Fleiss' Kappa values for the 30 items in CTTT ranged from .79 to .89, with an average value of .85. All items have values that are in the high category, indicating a very good level of inter-rater agreement. The high value of inter-rater reliability indicates

that the items in the CTTH were consistently rated as valid representations of the CT, thus strengthening the content quality of the instrument. Therefore, the instrument can effectively identify and evaluate students' CT in understanding the concepts of temperature and heat. Ultimately, this can provide meaningful feedback for teachers in designing more targeted learning strategies and encourage the strengthening of CT in physics learning in secondary schools.

4.3. Internal Consistency

Internal reliability is the main indicator to assess the consistency of the items in an instrument, which shows the extent to which the items together measure the intended construct. The internal consistency of the CTTH test was calculated using Cronbach's alpha coefficient, and a value of $\alpha = .86$ was obtained. The value is in the high category and is considered adequate for the context of educational instruments. According to Cohen (1988), alpha values above .70 are acceptable in educational research, while Penfield and Giacobbi (2004) recommend a minimum value of .80 for instruments used to evaluate psychological or cognitive abilities. Therefore, the value of $\alpha = .86$ indicates that the CTTH instrument has high internal consistency. The consistency indicates that the items in CTTH stably measure the critical thinking dimension on temperature and heat according to the six subscales formulated by Facione (2020). It is important in the context of learning because reliable instruments produce valid data to evaluate students' CT.

4.4. Item Difficulty

Item difficulty reflects how challenging a test item is, expressed in logits—positive values indicate harder items, negative values easier ones. Based on Table 4, CTTH items range from -0.82 to 1.84 logits, with item 11 being the most difficult and item 17 the easiest. This spread indicates that the CTTH instrument encompasses a broad range of cognitive complexity, enabling it to assess students across various ability levels. Such variation enhances its usefulness in evaluation, curriculum development, and targeted instructional interventions.

4.5. Item Discrimination

Item discrimination measures how well a test item differentiates between students with high and low critical thinking (CT) abilities. Values ranged from .23 to .64, with item 25 showing the highest discrimination (.64) and item 20 the lowest (.23). While item 20 value remains within the acceptable threshold, its relatively low discrimination suggests limited effectiveness in distinguishing student CT levels. Such limitations highlight the need for further review or revision to enhance its diagnostic utility. Despite this, the overall distribution confirms the CTTH instrument's capacity to differentiate CT abilities effectively, supporting its role in competency-based assessments.

4.6. Rasch Model Fit Analysis

The Rasch fit analysis evaluates whether item responses align with model expectations. Most CTTH items fall within acceptable Infit MNSQ (.94–1.11) and Outfit MNSQ (.71–1.06) ranges. For instance, item 25 shows a strong fit (Infit: 0.92, Outfit: 0.71), while item 11 remains within bounds (Infit: 1.05, Outfit: 1.06). These results affirm that CTTH items adequately conform to Rasch model assumptions, reinforcing the instrument's validity and reliability in assessing critical thinking in physics.

4.7. Measurement of CT components on the Topic of Temperature and Heat

The CTTH instrument successfully measures six components of critical thinking (CT) on the topic of temperature and heat: interpretation, analysis, evaluation, explanation, inference, and self-regulation (see Table 5). All components show difficulty levels within an appropriate range (from -0.5 to $.54$ logits) and acceptable Infit and Outfit MNSQ values, indicating good model fit. Interpretation (difficulty = -0.324 , Infit = .998) is relatively easy and shows strong model alignment. Analysis (-0.5 , Infit

= 1.028) also performs well, though slightly above the ideal MNSQ range, suggesting a few items might benefit from revision. Evaluation (–.12, Infit = 1.022) is accessible to students, but minor misfits may reflect shallow engagement with evaluative tasks. Explanation (.026, Infit = .992) and inference (.544, Infit = .96) are more cognitively demanding, indicating good discrimination of higher-level skills. Self-regulation (.38, Infit = 1.002) also fits well, highlighting the instrument's ability to capture metacognitive aspects. These findings support the use of CTTH as a reliable diagnostic tool for identifying student CT profiles and informing targeted instruction in physics education.

Table 5. Measurement Results of Each Component of Critical Thinking Skills

CT components	Correct answers	Difficulty	Standard Error	Infit MNSQ (logit)	Outfit MNSQ (logit)
Interpretation	145.6	-.324	.098	.998	.82
Analysis	161	-.5	.092	1.028	1.046
Evaluation	137	-.12	.112	1.022	1.058
Explanation	114.2	.026	.11	.992	.994
Inference	94.6	.544	.15	.96	.772
Self-regulation	96	.38	.126	1.002	1.03

5. Discussion

The study aims to develop and validate CTTH, an instrument specifically designed to measure CT in the context of temperature and heat topics in physics. The importance of CT in science is undeniable, given that this ability is essential for students to analyze, evaluate, and make inferences based on scientific data. However, many existing CT test instruments are not designed to measure CT in the specific context of certain science fields, such as physics. This shortcoming creates a gap that must be filled to support science learning more effectively. The development of the CTTH test involved an initial validation process that included expert reviews. The validation process aims to ensure that the instrument developed is of good quality in terms of validity and reliability. The initial validation results suggest that CTTH is a good fit with the Rasch model, and the items in the test can distinguish appropriately between students who have high and low CT. Therefore, CTTH is expected to be an effective tool to measure and enhance students' CT in physics learning, particularly on temperature and heat, and contribute to developing domain-specific CT test instruments in science.

Reliability findings confirmed that the Cronbach's alpha value for CTTH was .86, which indicates good and acceptable reliability in the context of educational instruments. It indicates that the CTTH test has a fairly high level of consistency, so the CTTH instrument can be relied upon to measure students' CT accurately. CTTH is designed to cover a range of complex and diverse cognitive skills, corresponding to the six CT subscales Facione (2020) described, such as interpretation, analysis, evaluation, inference, explanation, and self-regulation. Therefore, test reliability is very important, given the diverse types of skills being measured (Casella et al., 2020; Engelhard & Wang, 2021; Zehirlioglu & Mert, 2020). Furthermore, the characteristics of the participants in this study consisted of high school students who had completed learning about temperature and heat. The test was administered after the lesson to ensure the participants could apply the CT in a relevant context. Therefore, the reliability results show that CTTH can effectively assess students' CT on temperature and heat and provide a strong basis for its use in further research or learning.

The discriminative ability analysis of the CTH instrument indicates that although some items have a high difficulty level, they still have a good discrimination value. It indicates that the items can clearly distinguish between students who have high and low CT, thus remaining effective in measuring different levels of CT. Therefore, even if items are quite difficult, deleting them at this stage is not recommended, as deleting items may compromise the test's content validity (Affandy et al., 2021; Kaur et al., 2024). Each item contributed to the comprehensive measurement of CT, which was the main goal of the test development. The recommendation for further development is to conduct further validation using a larger and more diverse sample. It is important to ensure that the test can be applied to various student populations with different backgrounds and to test the consistency and validity of the test items in various contexts.

The interpretation component showed a high mean score of 145.6 with a relatively low difficulty level of $-.324$ logit. This indicates that students can easily solve items related to interpreting physics information, particularly on the topic of temperature and heat. Interpretation skills are fundamental in physics learning as they involve understanding data representations, graphs, and conceptual narratives of physical phenomena. Given the lower difficulty level, it suggests that most students have adequate cognitive skills to access and construct meaning from the provided information. These findings align with Kassiavera et al. (2024), who found that interpretation skills can be developed early through contextual learning and phenomenon-based problems. Similarly, Khoiriza et al. (2021) identified interpretation as one of the easiest aspects of critical thinking (CT) to enhance through explicit instruction. However, the possibility that the questions may be too simple and may not adequately differentiate higher levels of student ability should be considered. Nonetheless, the excellent Rasch model fit (Infit MNSQ = $.998$, Outfit MNSQ = $.82$) supports that the items remain valid for measuring interpretation skills.

The analysis component, with a mean score of 161 and a difficulty level of $-.5$ logits, posed more of a challenge for students. Analysis involves identifying relationships between concepts, interpreting scientific arguments, and evaluating causal elements in physical phenomena. These skills require higher-level cognitive abilities, logical thinking, and systematic problem-solving, which some students have not fully mastered (Danday & Monterola, 2019; Mundilarto & Ismoyo, 2017). Research by Facione (2020) confirms that analytical skills form the foundation of CT and develop more slowly, especially in environments where these skills are not explicitly trained. Furthermore, Mundilarto and Ismoyo (2017) suggest that analytical skills are strengthened by students' engagement in open-ended and investigative problems. Despite slight deviations in the Rasch model fit (Infit MNSQ = 1.028 , Outfit MNSQ = 1.046), the analysis component still operates within the acceptable limits, confirming that it validly measures students' analytical skills.

The evaluation component showed a mean score of 137 with a balanced difficulty level of -0.12 logits. The Rasch model fit values (Infit MNSQ = 1.022 , Outfit MNSQ = 1.058) indicate that the instrument effectively measures students' ability to assess the quality and credibility of information, particularly in the context of temperature and heat concepts. Evaluation skills are crucial in CT as they allow students to distinguish valid information from unreliable sources, identify underlying assumptions, and evaluate scientific arguments. Affandy et al. (2021) found that trained evaluation skills significantly enhance students' scientific problem-solving abilities. Facione (2020) also emphasized the importance of evaluation in logical, evidence-based decision-making. Although some students with limited analytical thinking skills may struggle with making accurate judgments, the fit indices show that the evaluation items remain reliable for assessing students' ability to evaluate scientific information.

The explanation component, with a mean score of 114.2 and a difficulty level of $.026$ logits, was relatively easier for students. This suggests that students were able to provide explanations of concepts related to temperature and heat, but perhaps with less depth or clarity. The ability to

explain is an essential aspect of CT, as it involves organizing and articulating scientific reasoning logically and coherently. Affandy et al. (2021) argue that good scientific explanations reflect deep conceptual understanding and contribute to forming strong scientific arguments. Facione (2020) also highlights the importance of explanation in demonstrating understanding and the ability to communicate it effectively. Despite the lower mean score, the excellent fit indices (Infit MNSQ = .992, Outfit MNSQ = .994) indicate that the explanation component reliably measures students' explanatory skills and can be used to guide teaching strategies focused on enhancing students' ability to explain scientific phenomena.

The inference (mean = 94.6, difficulty = .544 logits) and self-regulation (mean = 96, difficulty = .38 logits) components presented more challenges for students, suggesting that these metacognitive skills require greater effort to develop. Inference skills require students to draw logical conclusions from data, while self-regulation involves evaluating and controlling one's thinking and learning processes (Facione, 2000; Kim et al., 2025; Sarigoz, 2012). Both skills are essential in evidence-based decision-making and managing one's cognitive processes effectively. Research by Kassiavera et al. (2024) and Facione (2020) suggests that these skills develop more slowly and often require explicit instruction and continuous practice. Although the Rasch model fit values for these components (Infit MNSQ = .96 and 1.002; Outfit MNSQ = .772 and 1.03) show slight variation, they remain within acceptable limits, supporting the continued use of these components to measure CT skills in physics.

6. Conclusion

The current study successfully developed and validated the Critical Thinking Test in Temperature and Heat (CTTH) as an effective instrument to measure students' critical thinking skills on temperature and heat in physics. The validity and reliability of the CTTH instrument showed favourable validity with a Rasch model fit, indicating that the test items could distinguish between students with high and low levels of critical thinking skills. The instrument also had good reliability, with a Cronbach's alpha value of .72, indicating adequate internal consistency and a kappa value of .83 for inter-rater reliability, indicating high consistency between observers. The items on the CTTH exhibited good discrimination ability. Although some items have a high difficulty level, they can still distinguish students with high and low CT, making them effective in measuring various levels of CT ability. The CTTH instrument was successfully used to measure students' CT on six components, according to Facione (2020), which are interpretation, analysis, evaluation, inference, explanation, and self-regulation. Each component exhibited an appropriate difficulty level, with the interpretation component relatively easy and the inference and self-regulation components more challenging.

7. Suggestion

The CTTH instrument can be used effectively to measure CT in the context of physics, especially on the topic of temperature and heat, and contributes to the development of more specific CT assessment instruments in the field of science. The study also opens opportunities for further development in integrating CT in physics learning and provides further insights into its application in education. Future research is suggested to expand the critical thinking assessment framework and answer research questions for wider application in science education.

Declarations

Author Contributions. B.S.: Conceptualization, Methodology, Data Collection, Data Analysis, Writing – Original Draft Preparation. W.S.: Supervision, Validation, Review & Editing, Project Administration, Corresponding Author. B.A.P.: Instrument Development, Formal Analysis, Visualization, Writing – Review & Editing. N.Y.I.: Literature Review, Data Interpretation, Writing – Review & Editing, Validation. All authors have read and approved the publication of the final version of the article.

Conflicts of Interest. The authors declare no conflict of interest.

Ethical Approval. The author declares that this research has obtained ethical permission from the informants ((secondary school students) and parents)

Acknowledgments. The authors would like to express their sincere gratitude to Sebelas Maret University (UNS), Surakarta, Indonesia, for providing the necessary facilities and support throughout the research process. During the preparation of this work, the author(s) used DeepL and Connected Papers to facilitate the translation and enhance the conceptual mapping of related works. The author(s) declare that they reviewed and edited the final output as needed and take(s) full responsibility for the content of the published article.

References

- 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(1), 1–10. <https://doi.org/10.1088/1742-6596/1842/1/012032>
- Affandy, H., Sunarno, W., Suryana, R., & Harjana. (2024). Integrating creative pedagogy into problem-based learning: The effects on higher order thinking skills in science education. *Thinking Skills and Creativity*, 53, Article 101575. <https://doi.org/10.1016/j.tsc.2024.101575>
- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959. <https://doi.org/10.1177/001316448004000419>
- Alpizar, D., Vo, T., French, B. F., & Hand, B. (2022). Growth of critical thinking skills in middle school immersive science learning environments. *Thinking Skills and Creativity*, 46, Article 101192. <https://doi.org/10.1016/j.tsc.2022.101192>
- Alwan, A. A. (2011). Misconception of heat and temperature among physics students. *Procedia - Social and Behavioral Sciences*, 12, 600–614. <https://doi.org/10.1016/j.sbspro.2011.02.074>
- Arnold, M., & Millar, R. (1994). Children's and lay adults' views about thermal equilibrium. *International Journal of Science Education*, 16(4), 405–419. <https://doi.org/10.1080/0950069940160403>
- Bajracharya, R. R., Emigh, P. J., & Manogue, C. A. (2019). Students' strategies for solving a multirepresentational partial derivative problem in thermodynamics. *Physical Review Physics Education Research*, 15(2), Article 20124. <https://doi.org/10.1103/PhysRevPhysEducRes.15.020124>
- 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>
- Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (4th ed.). Routledge. <https://doi.org/10.4324/9781315814698>
- Cari, C., Pratiwi, S. N., Affandy, H., & Nugraha, D. A. (2020). Investigation of undergraduate student concept understanding on Hydrostatic Pressure using two-tier test. *Journal of Physics: Conference Series*, 1511(1). <https://doi.org/10.1088/1742-6596/1511/1/012085>
- Cascella, C., Giberti, C., & Bolondi, G. (2020). An analysis of Differential Item Functioning on INVALSI tests, designed to explore gender gap in mathematical tasks. *Studies in Educational Evaluation*, 64, Article 100819. <https://doi.org/10.1016/j.stueduc.2019.100819>
- Cetin-Dindar, A., & Geban, O. (2011). Development of a three-tier test to assess high school students' understanding of acids and bases. *Procedia - Social and Behavioral Sciences*, 15, 600–604. <https://doi.org/10.1016/j.sbspro.2011.03.147>

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates. <https://doi.org/10.4324/9780203771587>
- Collado, S., Evans, G. W., Corraliza, J. A., & Sorrel, M. A. (2015). The role played by age on children's pro-ecological behaviors: An exploratory analysis. *Journal of Environmental Psychology*, 44, 85–94. <https://doi.org/10.1016/j.jenvp.2015.09.006>
- Cvenic, M. K., Planinic, M., Susac, A., Ivanjek, L., Jelcic, K., & Hopf, M. (2022). Development and validation of the Conceptual Survey on Wave Optics. *Physical Review Physics Education Research*, 18(1), 10103. <https://doi.org/10.1103/PhysRevPhysEducRes.18.010103>
- Danday, B. A., & Monterola, S. L. C. (2019). Effects of microteaching multiple-representation physics lesson study on pre-service teachers' critical thinking. *Journal of Baltic Science Education*, 18(5), 692–707. <https://doi.org/10.33225/jbse/19.18.692>
- Engelhard, G., & Wang, J. (2021). *Rasch Models for Solving Measurement Problems: Invariant Measurement in the Social Sciences*. SAGE Publications. <https://doi.org/10.4135/9781071878675>
- Ennis, R. H. (1958). An appraisal of the watson-glaser critical thinking appraisal. *Journal of Educational Research*, 52(4), 155–158. <https://doi.org/10.1080/00220671.1958.10882558>
- 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>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–382. <https://doi.org/10.1037/h0031619>
- Hsu, F. H., Lin, I. H., Yeh, H. C., & Chen, N. S. (2022). Effect of Socratic Reflection Prompts via video-based learning system on elementary school students' critical thinking skills. *Computers and Education*, 183, Article 104497. <https://doi.org/10.1016/j.compedu.2022.104497>
- Kaltakci, D., Eryilmaz, A., & McDermott, L. C. (2016). Identifying pre-service physics teachers' misconceptions and conceptual difficulties about geometrical optics. *European Journal of Physics*, 37(4), Article 045705.
- Karaca-Atik, A., Meeuwisse, M., Gorgievski, M., & Smeets, G. (2023). Uncovering important 21st-century skills for sustainable career development of social sciences graduates: A systematic review. *Educational Research Review*, 39, Article 100528. <https://doi.org/10.1016/j.edurev.2023.100528>
- Kassiavera, S., Suparmi, A., Cari, C., & Sukarmin, S. (2024). Application of Rasch Model in two-tier test for assessing critical thinking in physics education. *Journal of Baltic Science Education*, 23(6), 1227–1242. <https://doi.org/10.33225/jbse/24.23.1227>
- Kaur, R., Mantri, A., Nagabhushan, P., & Singh, G. (2024). Rasch Computing Analysis of Two Tier Concept Inventory to Assess Engineering Students' Conceptual Knowledge. *SN Computer Science*, 5(5), 643–656. <https://doi.org/10.1007/s42979-024-02955-6>
- Khoiriza, I., Aminatun, T., Pramusinta, W., & Hujatulatif, A. (2021). Science learning and environment: Analysis of student's scientific literacy based on Indonesia's waste problem. *Proceedings of the 6th International Seminar on Science Education*, 541(Isse 2020), 775–779. <https://doi.org/10.2991/assehr.k.210326.111>
- Kim, L., Imjai, N., Kaewjomnong, A., Dowpiset, K., & Aujiapongpan, S. (2025). Does experiential learning matter to strategic intuition skills of MBA students? Implications of diagnostic

- capabilities and critical thinking skills. *International Journal of Management Education*, 23(2), Article 101138. <https://doi.org/10.1016/j.ijme.2025.101138>
- Kinoshita, H. (2022). Teaching of critical thinking skills by science teachers in Japanese primary schools. *Journal of Baltic Science Education*, 21(5), 801–816. <https://doi.org/10.33225/jbse/22.21.801>
- Lawson, A. E. (1992). What do tests of “formal” reasoning actually measure? *Journal of Research in Science Teaching*, 29(9), 965–983. <https://doi.org/10.1002/tea.3660290906>
- Leach, S. M., Immekus, J. C., French, B. F., & Hand, B. (2020). The factorial validity of the Cornell Critical Thinking Tests: A multi-analytic approach. *Thinking Skills and Creativity*, 37, Article 100676. <https://doi.org/10.1016/j.tsc.2020.100676>
- Mafinejad, M. K., Arabshahi, S. K. S., Monajemi, A., Jalili, M., Soltani, A., & Rasouli, J. (2017). Use of Multi-Response format test in the assessment of medical students’ critical thinking ability. *Journal of Clinical and Diagnostic Research*, 11(9), LC10–LC13. <https://doi.org/10.7860/JCDR/2017/24884.10607>
- Mundilarto, & Ismoyo, H. (2017). Effect of problem-based learning on improvement physics achievement and critical thinking of senior high school student. *Journal of Baltic Science Education*, 16(5), 761–779. <https://doi.org/10.33225/jbse/17.16.761>
- Nurhuda, T., Rusdiana, D., & Setiawan, W. (2017). Analyzing students’ level of understanding on Kinetic theory of gases. *Journal of Physics: Conference Series*, 812, 12105. <https://doi.org/10.1088/1742-6596/812/1/012105>
- Penfield, R. D., & Giacobbi, P. R. (2004). Applying a score confidence interval to Aiken’s item content-relevance index. *Measurement in Physical Education and Exercise Science*, 8(4), 213–225. https://doi.org/10.1207/s15327841mpee0804_3
- Ridho, A. (2018). Does Multidimensionality Cause DIF? *ANIMA Indonesian Psychological Journal*, 33(2), 125. <https://doi.org/10.24123/aipj.v33i2.1583>
- Sapia, P., Napoli, F., & Bozzo, G. (2022). The Lawson’s test for scientific reasoning as a predictor for University formative success: A prospective study. *Education Sciences*, 12(11), 1–15. <https://doi.org/10.3390/educsci12110814>
- Sarigoz, O. (2012). Assessment of the high school students’ critical thinking skills. *Procedia - Social and Behavioral Sciences*, 46, 5315–5319. <https://doi.org/10.1016/j.sbspro.2012.06.430>
- Stolk, J. D., Gross, M. D., & Zastavker, Y. V. (2021). Motivation, pedagogy, and gender: examining the multifaceted and dynamic situational responses of women and men in college STEM courses. *International Journal of STEM Education*, 8, 1–19. <https://doi.org/10.1186/s40594-021-00283-2>
- Suwita, S., Saputro, S., Sajidan, S., & Sutarno, S. (2024). Assessing lower-secondary school students’ critical thinking skills in photosynthesis: A Rasch Model approach. *Journal of Baltic Science Education*, 23(6), 1278–1290. <https://doi.org/10.33225/jbse/24.23.1278>
- Svedholm-Häkkinen, A. M., Forzani, E., Coiro, J., & Kiili, C. (2025). Online credibility evaluation skills in upper secondary students: The role of grade level, argument evaluation, and analytic thinking dispositions. *Learning and Individual Differences*, 118, Article 102640. <https://doi.org/10.1016/j.lindif.2025.102640>
- Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students’ misconceptions in science. *International Journal of Science Education*, 10(2), 159–169.

<https://doi.org/10.1080/0950069880100204>

- Wind, S. A. (2019). Nonparametric Evidence of Validity, Reliability, and Fairness for Rater-Mediated Assessments: An Illustration Using Mokken Scale Analysis. *Journal of Educational Measurement*, 56(3), 478–504. <https://doi.org/10.1111/jedem.12222>
- Zakwandi, R., Istiyono, E., & Dwandaru, W. S. B. (2024). A two-tier computerized adaptive test to measure student computational thinking skills. *Education and Information Technologies*, 29(7), 8579–8608. <https://doi.org/10.1007/s10639-023-12093-w>
- Zehirlioglu, L., & Mert, H. (2020). Validity and reliability of the Heart Disease Fact Questionnaire (HDFQ): a Rasch measurement model approach. *Primary Care Diabetes*, 14(2), 154–160. <https://doi.org/10.1016/j.pcd.2019.06.006>

About the Contributor(s)

Bangun Sartono, is a doctoral student at the Doctorate Program of Natural Science Education, Universitas Sebelas Maret, Surakarta, Central Java, Indonesia. His research interests focus on science education, especially in the development and validation of assessment instruments and higher-order thinking skills in secondary education. He actively engages in interdisciplinary studies that integrate educational theory with empirical research in physics teaching and learning.

E-mail: bangunsartono@student.uns.ac.id

Widha Sunarno, is a professor in the Department of Physics Education at Universitas Sebelas Maret, Surakarta, Indonesia. He has extensive experience in the field of physics education, with a strong focus on instructional design, scientific literacy, and the integration of innovative teaching strategies to enhance students' conceptual understanding. As a senior researcher, he has supervised numerous postgraduate theses and published widely in both national and international journals.

E-mail: widhasunarno@staff.uns.ac.id

ORCID: <https://orcid.org/0000-0002-0805-8042>

Baskoro Adi Prayitno, is a lecturer in the Department of Biology Education at Universitas Sebelas Maret, Surakarta, Indonesia. His academic work centers on biology education, curriculum development, and the integration of technology in science teaching. He is actively involved in educational research and innovation, aiming to improve student engagement and scientific thinking through context-based learning approaches.

E-mail: baskoro_ap@fkip.uns.ac.id

ORCID: <https://orcid.org/0000-0002-1680-683X>

Nurma Yunita Indriyanti, is a lecturer in the Department of Chemistry Education at Universitas Sebelas Maret, Surakarta, Indonesia. Her areas of expertise include chemistry education, environmental chemistry, and the development of innovative teaching materials to foster critical and creative thinking in science learning. She is also an active contributor to various national and international academic forums related to science education.

E-mail: nurma.indriyanti@staff.uns.ac.id

ORCID: <https://orcid.org/0000-0002-9113-0235>

Publisher's Note: The opinions, statements, and data presented in all publications are solely those of the individual author(s) and contributors and do not reflect the views of Universitepark, EDUPIJ, and/or the editor(s). Universitepark, the Journal, and/or the editor(s) accept no responsibility for any harm or damage to persons or property arising from the use of ideas, methods, instructions, or products mentioned in the content.
