Psychometric validation on the heat and temperature-critical thinking skills test (HnT-CTST) using the rasch model

Raden Hanna Rifani[1], Duden Saepuzaman[1], Winny Liliawati[1]*, Lina Aviyanti[1], John Richard Huwae[1], Muhammad Zahran[2], Asep Sopandi[3]

[1]Department of Physics Education, Universitas Pendidikan Indonesia, Dr. Setiabudi Street, No.229 Bandung, 4015, Indonesia

[2]Graduate School of Education, Okayama University, Okayama, Japan

[3]Junior High School One of Cianjur, Siliwangi Street, No.94, Cianjur, Indonesia

*Corresponding author, email: winny@upi.edu

Article History
Received: 15 December 2025
Revised: 17 May 2026
Accepted: 30 July 2026

Keywords
Critical thinking
Heat and temperature
Rasch model
Two-tier test

Abstract
Critical thinking is an essential competency in physics learning, particularly in temperature and heat topics, which are often associated with misconceptions and difficulties in conceptual understanding. Therefore, a valid and reliable instrument is required to accurately assess students' critical thinking skills. This study aimed to develop and validate the Heat and Temperature–Critical Thinking Skills Test (HnT-CTST), a two-tier multiple-choice instrument constructed based on Facione's critical thinking framework. A quantitative approach with a cross-sectional survey design was employed. The instrument consisted of 18 items that were reviewed by five experts and administered to 88 eighth-grade students at a public junior high school in Cianjur Regency, Indonesia. Rasch analysis was used to evaluate the instrument's psychometric properties, including content validity, reliability, model fit, unidimensionality, item difficulty, and gender fairness through Differential Item Functioning (DIF) analysis. The content validity results showed Aiken's V values ranging from 0.96 to 1.00, indicating excellent validity. Item reliability reached 0.94, person reliability was 0.71, and KR-20 was 0.74, demonstrating satisfactory internal consistency. Infit and outfit MNSQ values for both persons and items were within the acceptable range, confirming good model fit. Principal Component Analysis (PCA) of residuals revealed that 37.4% of the variance was explained by the primary construct, supporting unidimensionality. The Wright Map indicated that the instrument effectively represented varying levels of student ability. DIF analysis found no significant gender bias across all items. These findings demonstrate that the HnT-CTST is a valid, reliable, and fair instrument for measuring junior high school students' critical thinking skills in temperature and heat concepts and can support future assessment and instructional practices.

Rifani, R, H., Saepuzaman, D., Liliawati, W., Aviyanti, L., Huwae, J, R., Zahran, M., & Sopandi A. (2026). Psychometric validation on the heat and temperature-critical thinking skills test (HnT-CTST) using the rasch model. *Momentum: Physics Education Journal*, *10*(1), 150-161. https://doi.org/10.21067/mpej.v10i1.13481

1. Introduction

Critical thinking is generally measured as the ability to reason clearly and rationally (Song & Cai, 2024; Song & Cai, 2024). The development of critical thinking (CT) skills is widely recognized as a primary goal in science education (Bailin, 2002). CT encompasses the ability to analyze, evaluate, and reflect on information to make logical decisions in solving scientific problems and foster lifelong scientific literacy (García-Carmona, 2025). To equip young people to face the dynamics of modern society and the challenges of increasingly complex scientific methods, many countries have initiated educational reforms by making CT integration into science curricula a top priority in recent decades (Ramadani et al., 2025). In Indonesia, CT has been designated as one of the core competencies that need to be developed at the junior high school level. This is reflected in the direction of the national curriculum policy, which emphasizes the importance of CT integration in the science learning process as part of the expected learning outcomes.

Science learning, particularly physics, presents unique challenges in developing conceptual understanding. Students often struggle to connect abstract concepts such as temperature and heat to real-world phenomena (Salame et al., 2025). Misunderstandings of the relationship between heat and temperature, or misunderstandings about the process of heat energy transfer within an object,

doi: 10.21067/mpej.v10i1.13481
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2548-9135

are common and can hinder deeper conceptual understanding ((Brundage et al., 2024; Tugirin et al., 2025; Utama & Toifur, 2025). These emerging challenges in education highlight the urgency of designing innovative learning strategies, along with assessment instruments that go beyond mere mastery of the material. Assessments need to be designed to comprehensively measure reasoning capacity and problem-solving skills, thus truly reflecting student competencies relevant to the demands of the times.

To address these needs, it must be recognized that ordinary multiple-choice questions are insufficient to measure critical thinking on abstract topics like temperature and heat. Traditional multiple-choice format only capture the final answer, which fails to reveal the students' underlying scientific reasoning or differentiate between genuine understanding and mere guessing (Caleon & Subramaniam, 2010; Gurel et al., 2015; Ku, 2009). Consequently, much literature utilizes innovative assessment models (Mukhlisa Nurul, 2025; Nurazizah Aliah et al., 2023). In line with the development of the paradigm assessment as learning, selecting evaluation instruments that can reflect and simultaneously encourage higher-order thinking skills is very important. Within this framework, two-tier test are developed and validated as a tool capable of providing qualitative information on students' CT (Maulidiansyah et al., 2021). This model test is not only a measuring tool but also a source of information for learning reflection. In the model questions two-tier. The first level typically assesses factual, conceptual, or procedural knowledge. Meanwhile, the second level asks students to justify their answers at the first level, thus demonstrating the depth of students' CT (Lestari et al., 2023).

Historically, many previous instruments evaluating critical thingking still reliead in Classical Test Theory (CTT). However, CTT has significant inherent limitations, particularly its sample and test dependency, meaning the item rather than being objective measurements(Boone, 2016; Sahin & Anil, 2017). To overcome these limitations, Rasch modeling is used as a measurement model that can provide detailed information regarding the quality of items in the instrument used (Sunni et al., 2024; Sunni et al., 2024). By analyzing the level of item difficulty, respondent ability, and response patterns, Rasch analysis provides evidence of validity and reliability, and ensures that the instrument measures the intended construct consistently across various student populations(Bond Trevor G & Fox Christine M, 2012; Kassiavera et al., 2024). Several recent studies have successfully applied Rasch modeling to two-tier tests in Physics Education, and shown that this model is effective in measuring higher-order thinking skills (Kassiavera et al., 2024; Kusumadani et al., 2025; Zivanka et al., 2022).

Despite these advancements, there remains a lack of objectively validated tools tailored specifically for complex thermodynamic concepts. Therefore, this study aims to fill the gap by developing an objective, Rasch-validated critical thinking instrument specifically for Heat and Temperature. In line with these developments, this study reports the design and validation of an instrument. Two-tier test specifically aimed at measuring critical thinking skills on temperature and heat material at the junior high school level. This instrument was analyzed using Rasch modeling to evaluate its psychometric characteristics. This study was formulated to answer: (1) what is the level of validity and reliability of the critical thinking test instrument with a two-tier multiple-choice model based on Rasch modeling in measuring students' understanding of temperature and heat material?; (2) Does the DIF analysis identify any gender bias in the instrument items used? From these research questions, researchers can ensure the feasibility of the instrument as a reliable measuring tool in secondary physics education. This study contributes to the development of CT assessment research and offers educators a validated assessment framework to support the development of higher-order thinking skills in students on the topic of temperature and heat.

2. Method

This study employed a quantitative approach by applying a cross-sectional survey model. This approach was chosen to enable accurate numerical measurements and statistical analysis that yield reliable and accurate conclusions that can describe students' Critical Thinking (CT) levels and test the instrument's psychometric quality. The focus of the study was to validate the Heat and Temperature-Critical Thinking Skills Test (HnT-CTST) through Rasch modeling analysis.

The sampling method used in this study was purposive sampling, which was applied to obtain participants who met specific criteria. The sample consisted of 88 students, with a total of 42 male participants and 46 female students, at a public junior high school in Cianjur Regency, West Java. The

sample size was considered moderate, meeting the recommended guidelines of 50 to 100 participants to achieve stable item calibration in Rasch analysis when the instrument has fewer than 20 items (Linacre, 1994). Furthermore, at the time of the test, all participants had completed learning about temperature and heat. This background understanding enabled them to provide more in-depth responses and minimize inaccurate or guesswork. Given that the primary objective of this study was to validate the psychometrics of the instrument, not to generalize the findings, the use of a targeted and homogeneous sample was deemed methodologically appropriate. The Rasch model emphasizes the quality of the interaction between the items and the respondents rather than simply sample size, and previous research has supported the reliability of parameter estimates obtained from a homogeneous academic group (Bond Trevor G & Fox Christine M, 2012; Boonee· et al., 2014). The students who were the research sample had completed and passed the Temperature and Heat learning chapter. This criterion was determined to ensure that the research sample had an adequate and relevant knowledge base, so that their responses to the HnT-CTST instrument could reflect the Critical Thinking abilities being measured.

Data for the CT test was collected using an online form. Participants were already accustomed to using digital platforms, and the online test was chosen for its time efficiency and feasibility without compromising the quality of the measurement instrument. Participants only took the test once to ensure the security and confidentiality of the test items. The test was administered within 60 minutes prior to administration to maintain participant consistency and concentration.

The Heat and Temperature-Critical Thinking Skills Test (HnT-CTST) is designed as a special assessment tool to measure students' Critical Thinking abilities on the subject of Temperature and Heat. This material was chosen because it is one of the materials that always appears in the national student olympiad, and is one of the materials that is considered difficult based on the results of student learning outcomes at the end of the semester at the school used as the instrument trial site. This test is structured in a two-tier multiple-choice format consisting of 18 carefully constructed questions, each aligned with Critical Thinking skills as described in the (Facione, 2020) framework. Facione's framework was selected over alternatives like Bloom's Taxonomy or Ennis's model because it provides a more highly operationalized set of skills-sucs as interpretating, analysis, and explanation-that direcly align with the cognitive demands of solving physics problems and justifying reasoning in a two-tier format. While Bloom's taxonomy is often too broad and hierarchical, and Ennis's framework focuses heavily on general logical argumentation, Facione's model specifivallly facilitates the measurements of how students interpret physical phenomenon and logically explain the underlying thermodynamic mechanisms (Giri & Paily, 2020). This design ensures that the instrument not only assesses content knowledge, but also digs deeper into how participants can analyze, interpret, and justify their answers, thus making it a comprehensive measuring tool for higher-order thinking skills. The HnT-CTST instrument goes through an expert judgment validation process before being used, to ensure its conformity with the Critical Thinking framework from Facione (2020). This test was developed on six core CT skills adapted to the topic of Temperature and Heat, along with its sub-materials as explained in Table 2. Meanwhile, examples of test items are shown in Table 1.

Table 1. Examples of CT Question Items in Skill Evaluation

Indicator Item	Question Form
It is interesting to conclude the results of the temperature scale conversion with actual conditions.	James is Andi's friend from the United States. He explained that in New York City, the average summer temperature reaches 84.2°F. Andi was surprised, thinking that was extremely hot. James then explained that the temperature scale used in America is Fahrenheit, while in Indonesia it is Celsius. Tier-one Questions Based on this information, what conclusions can be made about the differences between the Fahrenheit scale and the Celsius scale? The value on the Fahrenheit scale is always greater than on the Celsius scale for the same temperature. Values on the Fahrenheit scale are always smaller than those on the Celsius scale. Both scales show the same value even though they measure different objects. The Fahrenheit scale is used to measure low temperatures while Celsius is used for high temperatures. Tier-two Questions Based on the first-level question, how should Andi interpret the temperature of 84.2°F mentioned by James? Summer temperatures in New York are much hotter than in Indonesia.

Indicator Item	Question Form
	Summer temperatures in New York are actually almost the same as the average temperature in Indonesia. The Fahrenheit scale is lower than Celsius so the numbers look smaller. The temperature in New York cannot be compared with Indonesia because of the difference in latitude.

Table 2. Relationship between Critical Thinking Skills and the Topic of Temperature and Heat

Core Skill	Sub-Skill	Sub Topic		
		Temperature	Heat	Expansion
Interpretation	Categorize, understand important meanings, and explain meanings	S3	S1, S2, and S3	
Analysis	Reviewing ideas, identifying ideas, and analyzing arguments		S4 and S6	S5
Inference	Assessing claims and evaluating arguments	S8 and S9	S7, S8 and S9	
Evaluation	Review evidence, weigh alternatives, and draw conclusions	S10 and S12	S10 and S11	
Explanation	Presenting results, justifying procedures, and presenting arguments	S14 and S15	S14 and S15	S13 and S14
Self-regulation	Check yourself, and improve yourself	S17	S16, S17 and S18	

After going through the validation and revision process, the instrument was tested in a field study conducted in October 2025. The results of the V-Aiken analysis on each question item on the HnT-CTST instrument are presented in Table 3.

Table 3. Aiken V Item Analysis

No	Rate	Level of confidence (1%)
1	0.97	Valid
2	1.00	Valid
3	0.98	Valid
4	0.96	Valid
5	0.99	Valid
6	1.00	Valid
7	0.99	Valid
8	0.99	Valid
9	0.99	Valid
10	1.00	Valid
11	0.99	Valid
12	0.99	Valid
13	0.99	Valid
14	0.99	Valid
15	0.97	Valid
16	0.97	Valid
17	1.00	Valid
18	1.00	Valid

Based on Table 1. Aiken's V analysis of the test items, the average validity value for 18 test items ranged from 0.96 to 1.00. All test items were categorized as valid at a 99% level of confidence (1%). All test items were declared valid because the Aiken's V value was above the minimum validity of 0.93 (Hidayah1, 2023; Nurjanah et al., 2023). The HnT-CTST instrument was declared valid and could be tested.

The data analysis in this study used Rasch modeling to evaluate the psychometric characteristics of the instrument used. Reliability evaluation was seen based on person data and item reliability as well as CR Alpha (KR-20). Instrument validity was carried out by examining several statistical indicators, such as infit statistics and Outfit Mean Square (MNSQ) which refers to the tolerance limit of 0.5 to 1.5. Then the infit and outfit ZSTD from the data resulting from the Rasch modeling analysis, as well as the construct validity obtained. As further confirmation, the unidimensional test applied was Principal Component Analysis (PCA) on the Rasch model residuals. The results of this analysis ensured that the developed instrument truly measures a single construct that was intended to be understood. Further analysis of Differential Item Functioning (DIF) to

determine whether the developed instrument had a bias towards gender, was also carried out. All psychometric analyses mentioned above will be discussed one by one in the Results and Discussion section.

3. Results and Discussion

To support the validity and reliability of the data obtained, the HnT-CTST instrument underwent a series of psychometric tests. These tests provide an empirical basis for assessing the instrument's suitability for the construct being studied. As a summary of the main psychometric characteristics analyzed, Table 4 systematically presents the data to provide a clear picture of the instrument's quality.

Table 4. Summary of Psychometric Properties of the HnT-CTST Instrument

Category	Value
Number of Students	88
Number of Items	18
Person Reliability	.71
Item Reliability	.94
Cronbach's Alpha (KR-20)	.74
Infit MNSQ (Mean) Person	1.01
Outfit MNSQ (Mean) Person	.84
Infit MNSQ (Mean) Item	1.02
Outfit MNSQ (Mean) Item	.84
Infit ZSTD(Mean) Person	.11
Outfit ZSTD (Mean) Person	.06
Infit ZSTD (Mean) Item	.16
Outfit ZSTD (Mean) Item	-.20

Analysis of the HnT-CTST instrument shows that measurements using Rasch modeling conducted on 88 students with 18 items resulted in high item reliability of 0.94 and moderate person reliability of 0.71. High item reliability indicates the stability of the item difficulty parameters is very stable and consistent. Item reliability approaching 1.00 indicates that the instrument has a strong item construct and is not easily changed when tested on different groups of respondents (Bond Trevor G & Fox Christine M, 2012; Taber, 2018). With a value this high, it can be said that the difficulty ranking between items (from easiest to most difficult) can be maintained consistently. This stability is very important in Rasch modeling-based instruments because it ensures that each item functions according to its design and makes an appropriate contribution to the measurement of the construct (Kurniawan et al., 2024).

A person reliability value of 0.71 indicates that this instrument has a good ability to differentiate participants based on their ability levels. In the context of Rasch modeling, a person reliability value in the range of 0.70-0.79 is categorized as quite good (Frost et al., 2007). This value can be interpreted as indicating that student response patterns are relatively consistent across difficulty levels (Romdlon Maslahul Adi et al., 2022). In other words, the instrument is sensitive enough to separate students into several ability groups. If the instrument has low person reliability, conclusions about differences in ability between students will be less stable. However, in this study, a value of 0.71 indicates that the measurement results between individuals are reliable.

The KR-20 value obtained at 0.74 confirms that the instrument has good internal consistency, where each test item measures the same construct, the variance between items is relatively small, and there are no items that deviate from the measurement design (A Uyanah & U. I, 2023; Hari Sugiharto Setyaedhi, 2024; Stephen O. Ekolu & Harry Quainoo, 2019). This consistency is in line with the person and item reliability that have been obtained previously, so it can be concluded that the instrument has high reliability and is able to provide stable and reliable measurement results in the context of Educatio.

The infit and outfit Mean Square (MNSQ) indices are important indicators in Rasch modeling to assess the extent to which empirical data conform to model predictions, with ideal criteria ranging from 0.50 to 1.50 (Boonee· et al., 2014). The analysis results show that the infit value of the person MNSQ (Mean) of 1.01 and the outfit MNSQ (Mean) of 0.84 are very close to the ideal value of 1.00, so that the student response pattern can be said to be consistent with the model predictions without

many indications of random answers or extreme deviations. An infit value close to 1.00 indicates that students' responses to items with a level of difficulty commensurate with their abilities are regular, while an outfit value below 1.00 indicates there is no extreme outline in the response pattern (Sumintono Bambang & Widhiarso Wahyu, 2014). On the item side, the MNSQ infit value (mean) is 1.02 and the MNSQ outfit (mean) is 0.84, this value is also in the very good suitability range, indicating that all items function almost perfectly according to the Rasch modeling predictions and do not show deviant response patterns (Bond Trevor G & Fox Christine M, 2012)(Bond Trevor G & Fox Christine M, 2012).

A ZSTD value close to zero for both respondents and items indicates that the instrument has a consistent and stable response pattern, with no significant deviations from the expected distribution. In Rasch modeling analysis, ZSTD is a statistic that standardizes the Mean Square (MNSQ) value into a z-score with an acceptance range of -2 to +2. The analysis results show that the infit ZSTD (Mean) person = 0.11 and Outfit ZSTD (Mean) = -0.20, which indicates that all items behave normally, none are too misfit or overfit. This very small value close to zero is a previous finding from the MNSQ analysis that the instrument used is stable and valid to measure the intended construct. This finding is in line with previous research that used Rasch modeling to evaluate the quality of educational instruments, which confirmed that ZSTD values in the range of -2 to +2 indicate item and respondent consistency and are an indicator of instrument validity(Romdlon Maslahul Adi et al., 2022). Other research also shows that Rasch modeling analysis can accurately identify the quality of cognitive instruments, with ZSTD values close to zero as evidence of instrument stability (Sarifah et al., 2024). Furthermore, other research confirms that item fit statistics, such as ZSTD, are reliable indicators of model-data fit, as long as they are within a specified range (Müller, 2020). Thus, the results of the ZSTD analysis on the HnT-CTST instrument provide strong evidence that the instrument meets the psychometric criteria required by the Rasch approach.

Another noteworthy analysis is the Point Measure Correlation (PTMEA CORR), which is used to assess the consistency of the item's contribution to the measured construct. A positive correlation value is considered an indicator that the item provides relevant information regarding the participant's abilities (Mohamad et al., 2015). The analysis results show that the PTMEA CORR for all items ranges from 0.15 to 0.68, all positive, so that no item deviates from the construct's direction. This confirms that each item consistently contributes to the measurement, even though the correlation varies. This finding aligns with previous research, indicating that a positive correlation value in the PTMEA CORR indicates that the item functions in accordance with the measured construct, while a negative value can identify items that are misdirected or irrelevant (Azizah et al., 2022). Furthermore, other research emphasizes that the PTMEA CORR is an important indicator for verifying whether participant response patterns are consistent with Rasch model predictions (Müller, 2020). Thus, the results of the PTMEA CORR analysis on the HnT-CTST instrument provide strong evidence that all items have a consistent direction of contribution to the construct, so that this instrument can be considered valid and reliable.

The results of the unidimensionality analysis on the Hnt-CTST show that the instrument contains only one main construct with the results of the Rasch modeling explaining 37.4% of the total variance (raw variance explained by measures) as shown in Table 5. The value obtained far exceeds the minimum threshold of around 20%, which indicates the dominance of one dimension of the construct (Bond Trevor G & Fox Christine M, 2012).

Table 5. Principal Component Analysis of Rasch Residuals

Component	Explained Variance (%)
Raw Variance Explained by Measures	37.4
Unexplained First Contrast	8.6
Unexplained Second Contrast	7.6

Furthermore, the unexplained variance in the first contrast was only 8.6%, far from the critical lower limit of 15%, indicating that no significant secondary latent dimensions emerged. The eigenvalue for the first contrast was 2.48, which is still below the limit generally used to detect potential additional dimensions (Linacre, 1994). Thus, these results consistently support that the instrument is unidimensional, meaning that all items comprehensively measure one variable, namely CT.

In Rasch modeling, measure scores indicate the item's difficulty level. The ideal range typically used as a reference to describe a balanced item difficulty distribution is -2 < Measure < +2 logit. Items above 2 logit are considered too difficult, while items below -2 logit are considered too easy (Stemler & Naples, 2021). Based on the data obtained, the majority of items (14 of 18) are in the -2 to +2 logit range, so the difficulty level is appropriate to measure respondents' abilities representatively. Two items are too difficult, namely items 18 and 10, which indicates that only a few participants with high abilities are able to answer them. Meanwhile, two items are too easy, namely items 17 and 4, so they do not provide optimal discriminatory power for participants with low abilities. With this distribution, in general, the instrument is able to measure the range of abilities, but revisions may be needed for extreme items.

In Rasch modeling, the standard error of measurement (SEM) is used to describe the level of item accuracy in measuring respondents' abilities(Bond Trevor G & Fox Christine M, 2012). In general, the category used in SEM <0.5 logit is considered good, because it indicates high accuracy. Meanwhile, SEM >1.0 logit is considered inaccurate (Ekstrand et al., 2022; Linacre, 1994). Based on the analysis results obtained from the HnT-CTST instrument, all items have SEM values ranging from 0.25-0.31 logit, with an average of 0.36 logit and a standard deviation of 0.18 logit. This means that all items are in the good category, so there are no inaccurate items found in measuring CT. The relatively homogeneous consistency of SEM values also indicates that the estimated item difficulty is stable across all respondents' abilities. These findings align with research conducted by Huebner et al., (2021), which emphasized that SEM in Rasch can be used to assess measurement reliability, where low values indicate high estimation accuracy. Furthermore, Ekstrand et al. (2022) revealed that the logit transformation in Rasch maintains the linear nature of the measurement, making SEM an important indicator in ensuring instrument quality. Therefore, the HnT-CTST instrument can be said to have good and consistent measurement accuracy. The details of this data can be seen in Figure 1 below.

ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
18	16	75	2.85	.33	1.01	.10	1.14	.50	.47	.49	81.3	82.3	S18
10	21	75	2.35	.30	1.25	1.50	1.08	.38	.40	.52	72.0	79.6	S10
1	27	75	1.84	.28	1.15	1.08	1.19	.95	.44	.53	73.3	76.0	S1
12	34	75	1.29	.27	1.22	1.68	1.28	1.51	.39	.53	64.0	73.6	S12
17	35	75	1.22	.27	.94	-.48	.88	-.66	.57	.52	73.3	73.5	S17
9	46	75	.41	.27	.82	-1.66	.71	-1.33	.60	.48	80.0	72.6	S9
16	46	75	.41	.27	1.28	2.30	1.10	.49	.33	.48	58.7	72.6	S16
13	47	75	.34	.27	1.11	.94	.93	-.24	.43	.47	62.7	72.8	S13
8	48	75	.26	.27	.80	-1.82	.73	-1.13	.59	.47	82.7	73.0	S8
11	50	75	.11	.28	.87	-1.11	.74	-.99	.54	.45	80.0	73.2	S11
15	52	75	-.04	.28	1.20	1.68	1.22	.79	.31	.44	65.3	73.9	S15
14	54	75	-.20	.29	1.09	.78	.92	-.14	.38	.42	68.0	74.6	S14
6	55	75	-.29	.29	.84	-1.32	.63	-1.16	.53	.41	76.0	75.2	S6
7	55	75	-.29	.29	.85	-1.18	.62	-1.18	.53	.41	78.7	75.2	S7
2	69	75	-1.92	.44	.92	-.13	.48	-.54	.32	.23	92.0	92.0	S2
5	70	75	-2.13	.48	.96	.02	.54	-.37	.27	.21	93.3	93.4	S5
3	71	75	-2.38	.53	.99	.11	.59	-.24	.23	.19	94.7	94.7	S3
4	74	75	-3.85	1.02	.98	.31	.37	-.31	.15	.09	98.7	98.7	S4
MEAN	48.3	75.0	.00	.36	1.02	.16	.84	-.20			77.5	79.3	
P.SD	16.4	.0	1.66	.18	.15	1.21	.27	.80			11.4	8.7	

Figure 1. Measure Order

The following is Figure 2. This is the result of analysis using the Wright Map which shows that the instrument used has good capabilities in mapping variations in participant abilities.

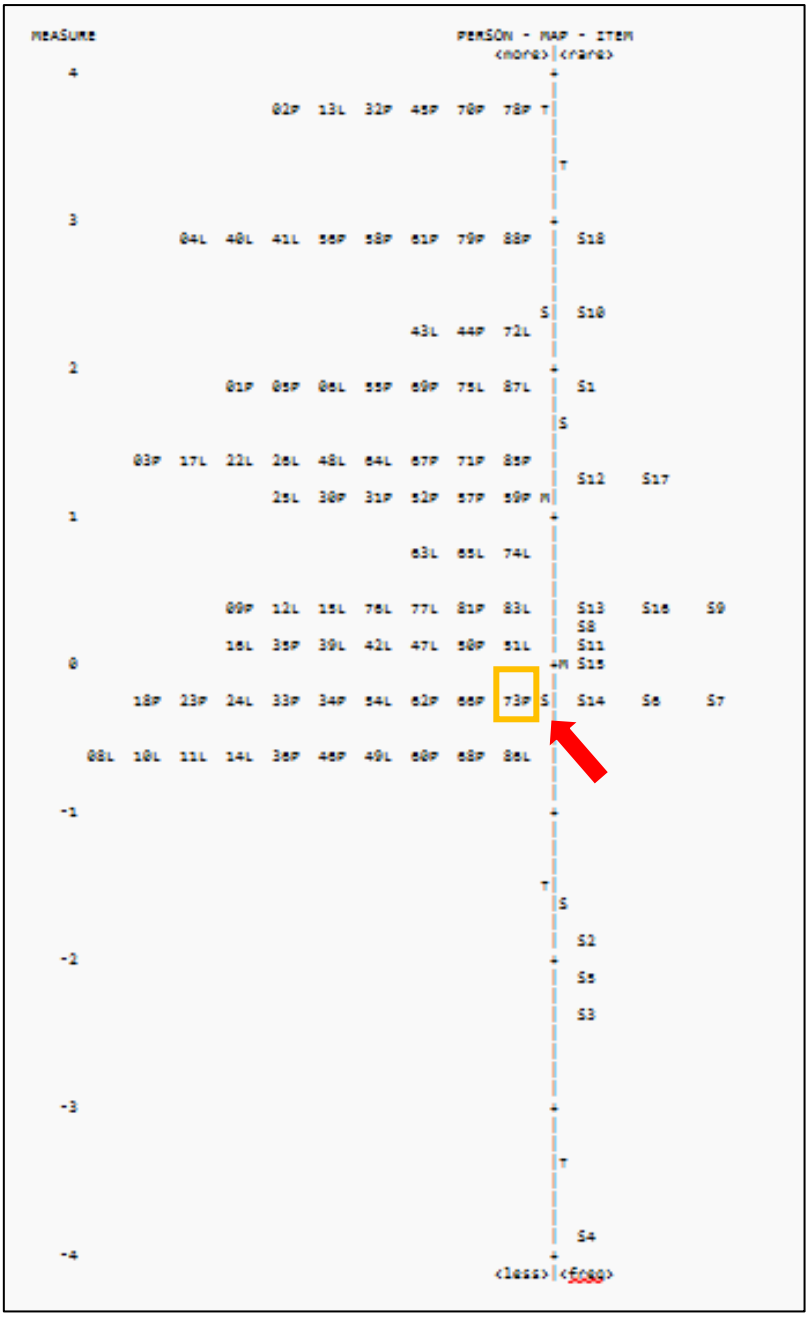


Figure 2. Wright Map

The distribution of items spanning a wide logit range demonstrates that the instrument was designed not only to assess basic abilities but also to capture the achievements of participants at higher levels of ability. The distribution of participants' abilities, concentrated in the middle to higher logit ranges, indicates that most participants were able to respond to items that required more complex analytical skills, thus indicating relatively good learning quality. Furthermore, the presence of items with a high level of difficulty is an added value because it allows researchers to identify participants who truly possess superior abilities. Let's take 73P as a reference. This person has an ability level equivalent to the difficulty level of items S15, S6, and S7. At this position, the person's probability of answering all three items correctly is 0.5. For items below the 73P person's position (which have a lower level of difficulty), the probability of answering correctly approaches 1.0. This indicates that the items below this position are relatively easier for this person. Conversely, for items located above his position (higher difficulty level), the probability of person 73P to answer correctly will be less than 0.5. This indicates that the questions above his position are a significant challenge for him. On the other hand, the diversity of item difficulty levels reflects that this instrument is not monotonous, but provides proportional challenges for each participant's ability range. Overall, the

Wright Map confirms that the instrument has a diverse and representative difficulty structure, but still requires the addition of several items at higher and lower difficulty levels to provide more balanced coverage and be able to capture participants' abilities more comprehensively.

To investigate the potential for gender bias affecting the difficulty level of a test item, a Differential Item Functioning (DIF) analysis was conducted. Figure 4 presents a visualization of these findings in the form of a person DIF plot, then **Figure** 3 shows the data used to analyze whether any bias arises due to gender.

PERSON CLASSES	SUMMARY DIF CHI-SQUARED	D.F.	PROB.	BETWEEN-CLASS/GROUP ITEM UNWTD MNSQ	ZSTD	Number	Name
2	4.1522	1	.0416	4.3807	1.82	1	S1
2	.5803	1	.4462	.5954	.13	2	S2
2	2.8931	1	.0890	4.0872	1.74	3	S3
2	.5004	1	.4793	.9515	.44	4	S4
2	.1401	1	.7082	.1425	-.54	5	S5
2	.1076	1	.7429	.1085	-.64	6	S6
2	.0617	1	.8038	.0631	-.81	7	S7
2	1.0891	1	.2967	1.1163	.55	8	S8
2	.3619	1	.5475	.3659	-.13	9	S9
2	.0990	1	.7530	.0993	-.67	10	S10
2	.3258	1	.5681	.3312	-.18	11	S11
2	.4445	1	.5049	.4502	-.02	12	S12
2	.5814	1	.4458	.5911	.13	13	S13
2	1.3394	1	.2471	1.3775	.71	14	S14
2	.0000	1	1.0000	.0001	-1.55	15	S15
2	2.3840	1	.1226	2.4754	1.22	16	S16
2	2.1327	1	.1442	2.2068	1.11	17	S17
2	1.3466	1	.2459	1.3841	.71	18	S18

Figure 3. Data of DIF

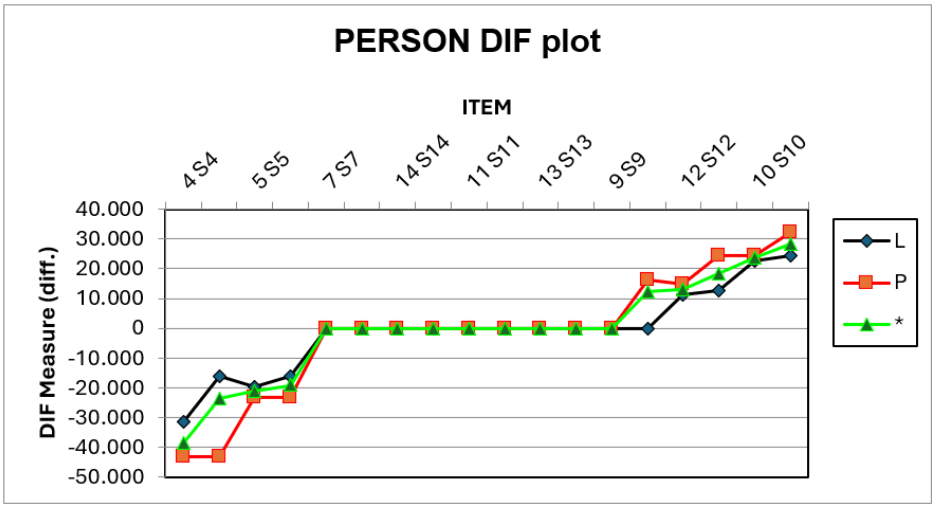


Figure 4. DIF Graph

The graph plots the comparison of DIF measurements between male (black line) and female (red line) respondent groups and the overall sample mean. The graph shows that the instrument performed consistently across male and female respondents. The line pattern in the graph shows that DIF values for both groups move parallel and close together across almost all items, with no extreme deviations indicating bias in favor of one group or the other. While some initial and final items do show slight differences between the two groups, these differences are within a reasonable range. Most items actually show overlap between males, females, and the overall mean, indicating that item difficulty levels are perceived relatively equally by both groups. Thus, these findings confirm that the instrument is free from significant gender bias and can be used fairly to assess the abilities of participants from both respondent groups.

Despite its strong psychometric performance, this study's generalizability is limited by a small, homogeneous sample, a specific focus on Temperature and Heat, and DIF analysis restricted to gender. Future research should validate this instrument using larger, diverse populations, expand it to other physics topics, and examine additional demographic variables to enhance its overall robustness and applicability.

4. Conclusion

This study successfully validated the HnT-CTST instrument as a measure of junior high school students' critical thinking skills on temperature and heat by applying Rasch modeling. A series of psychometric analyses, ranging from content validity through Aiken's V to evaluation of reliability, model fit, unidimensionality, and item difficulty distribution, demonstrated that this instrument has strong and consistent measurement quality. Very high item reliability and good internal consistency indicate that the items are able to measure the designed construct stably. Analysis of the level of item fit and responses also showed that student response patterns were in line with model predictions without any extreme deviations. Mapping through Wright Map confirmed that the instrument is able to capture variations in student abilities representatively, with item distributions covering a wide range of difficulty and relevant for analytical abilities. Furthermore, the results of the DIF examination showed that the instrument is free from gender bias, making it suitable for use as a fair measurement tool. Overall, this study confirms that the HnT-CTST is a valid, reliable, and devotion-based instrument that fulfills the objective measurement principles in assessing students' critical thinking skills in physics learning, especially in the material of Temperature and Heat.

Although this instrument has met good psychometric standards, several areas remain open for further research. The use of a sample from only one region requires caution in generalizing the results, and further studies with a wider range of schools and regional contexts are highly recommended. The current two-tier multiple-choice instrument could also be further developed by adding an open-response format or interactive technology-based tasks to capture a wider range of student reasoning skills. Furthermore, revisions to items that are either too easy or too difficult should be considered to ensure a more balanced distribution of difficulty and more optimally map student abilities. Utilizing a calibrated question bank for the development of adaptive assessments is also a potential direction that could improve measurement precision while providing a more relevant testing experience for students. With these developments, the HnT-CTST instrument is expected to continue to be refined into a more comprehensive and equitable evaluation tool that supports the quality of physics learning at the secondary level.

Author Contributions

All authors have equal contributions to the paper. All the authors have read and approved the final manuscript.

Funding

No funding support was received.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

A Uyanah, D., & U. I, N. (2023). The theoretical and empirical equivalence of Cronbach alpha and Kuder–Richardson Formula 20 reliability coefficients. International *Research Journal of Innovations in Engineering and Technology, 7*(5), 17–23. https://doi.org/10.47001/IRJIET/2023.705003

Azizah, N., Suseno, M., & Hayat, B. (2022). Item analysis of Rasch model items in the final semester examination of Indonesian language lessons. *World Journal of English Language, 12*(1), 15–26. https://doi.org/10.5430/wjel.v12n1p15

Bailin, S. (2002). Critical thinking and science education. *Science & Education, 11*(4), 361–375. https://doi.org/10.1023/A:1016042608621

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education, 15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Brundage, M. J., Meltzer, D. E., & Singh, C. (2024). Investigating introductory and advanced students' difficulties with change in internal energy, work, and heat transfer using a validated instrument. *Physical Review Physics Education Research, 20*(1), 010115. https://doi.org/10.1103/PhysRevPhysEducRes.20.010115

Caleon, I., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education, 32*(7), 939–961. https://doi.org/10.1080/09500690902890130

Ekstrand, J., Westergren, A., Årestedt, K., Hellström, A., & Hagell, P. (2022). Transformation of Rasch model logits for enhanced interpretability. *BMC Medical Research Methodology, 22*(1), Article 335. https://doi.org/10.1186/s12874-022-01816-1

Facione, P. A. (2020). *Critical thinking: What it is and why it counts (2020 update).* Insight Assessment.

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D., & Sloan, J. A. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health, 10*(Suppl. 2), S94–S105. https://doi.org/10.1111/j.1524-4733.2007.00272.x

García-Carmona, A. (2025). Scientific thinking and critical thinking in science education: Two distinct but symbiotically related intellectual processes. *Science & Education, 34*(1), 227–245. https://doi.org/10.1007/s11191-023-00460-5

Gurel, D. K., Eryılmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

Giri, V., & Paily, M. U. (2020). Effect of scientific argumentation on the development of critical thinking. *Science & Education, 29*(3), 673–690. https://doi.org/10.1007/s11191-020-00120-y

Setyaedhi, H. S. (2024). Comparative test of Cronbach's alpha reliability coefficient, KR-20, KR-21, and split-half method. *Journal of Education Research and Evaluation, 8*(1), 47–57. https://doi.org/10.23887/jere.v8i1.68164

Hidayah, N. (2023). Validity and reliability test of teaching materials using Aiken's V formula and SPSS 22. *Schola: Journal of Education and Learning, 1*(2), 62–68.

Huebner, A., & Skar, G. B. (2021). Conditional standard error of measurement: Classical test theory, generalizability theory and many-facet Rasch measurement with applications to writing assessment. *Practical Assessment, Research & Evaluation, 26*(1), Article 22. https://doi.org/10.7275/1ef7-rm16

Kassiavera, S., Suparmi, S., 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

Kurniawan, A., Istiyono, E., & Naba, S. D. (2024). Item quality analysis of physics concept understanding test with Rasch model. *Jurnal Ilmu Pendidikan Fisika, 9*(3), 474–484. https://doi.org/10.26737/jipf.v9i3.5692

Kusumadani, A. I., Afandy, H., Agustina, L., Astuti, R., & Waluyo, M. (2025). Evaluation of higher-order thinking skills of middle school students on vibration and wave topics using Rasch measurement. *Jurnal Penelitian Pendidikan IPA, 11*(5), 74–84. https://doi.org/10.29303/jppipa.v11i5.10900

Ku, K. Y. L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity, 4*(1), 70–76. https://doi.org/10.1016/j.tsc.2009.02.001

Lestari, N., Nisa, C., Mutazam, M., & Soeharto, S. (2023). Developing two-tier test of critical thinking ability using Rasch model. *Jurnal Ilmu Pendidikan Fisika, 8*(2), 259–268. https://doi.org/10.26737/jipf.v8i2.3737

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Maulidiansyah, D., Meutia, I., & Istiyono, E. (2021). Computer-based two-tier diagnostic test to identify critical thinking skills in optical instruments. *Jurnal Penelitian dan Evaluasi Pendidikan, 25*(2), 168–181.

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia - Social and Behavioral Sciences, 204,* 164–171. https://doi.org/10.1016/j.sbspro.2015.08.129

Müller, M. (2020). Item fit statistics for Rasch analysis: Can we trust them? *Journal of Statistical Distributions and Applications, 7*(1), Article 14. https://doi.org/10.1186/s40488-020-00108-7

Nurazizah, A., Natsir, N., & Zulkhaeriyah, Z. (2023). Development of innovative assessment methods in the context of modern learning. *International Journal of Education, Language, Literature, Arts, Culture, and Social Humanities, 1*(4), 84–93. https://doi.org/10.59024/ijellacush.v1i4.424

Nurjanah, S., Istiyono, E., Widihastuti, W., Iqbal, M., & Kamal, S. (2023). The application of Aiken's V method for evaluating the content validity of instruments that measure the implementation of formative assessments. *Journal of Research and Educational Research Evaluation, 12*(2), 125–135.

Ramadani, S., Sinaga, P., & Liliawati, W. (2025). Critical thinking in science learning: Systematic literature review. *Jurnal Pemberdayaan Masyarakat, 4*(1), 183–196. https://doi.org/10.46843/jpm.v4i1.378

Romadloni, N. M. A., Amaruddin, H., Adi, H. M. M., & Qurroti, I. A. L. (2022). Validity and reliability analysis using the Rasch model to measure the quality of mathematics test items of vocational high schools. *Journal of Research and Educational Research Evaluation, 11*(1), 93–103.

Salame, I. I., Fadipe, O., & Akter, S. (2025). Examining difficulties, challenges, and alternative conceptions students exhibit while learning about heat and temperature concepts. *Interdisciplinary Journal of Environmental and Science Education, 21*(2), e2510. https://doi.org/10.29333/ijese/15997

Sarifah, I., Kuswati, E., Suniasih, W., Fahrurrozi, F., & Susilo, G. (2024). Cognitive test instruments to measure student mathematics ability in elementary school: Rasch analysis. *Educational Administration: Theory and Practice, 30*(4), 1557–1568. https://doi.org/10.53555/kuey.v30i4.1557

Song, H., & Cai, L. (2024). Interactive learning environment as a source of critical thinking skills for college students. *BMC Medical Education, 24*(1), Article 683. https://doi.org/10.1186/s12909-024-05247-y

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. item response theory: Knowing when to cross the line. *Practical Assessment, Research & Evaluation, 26*(1), 1–16. https://doi.org/10.7275/v2gd-4441

Ekolu, S. O., & Quainoo, H. (2019). Reliability of assessments in engineering education using Cronbach's alpha, KR and split-half methods. *Global Journal of Engineering Education, 21*(1), 24–29.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial (Edisi revisi).* Trim Komunikata Publishing House.

Sunni, F. M., Dawana, I. R., Arrahmaniyyah, R., Sunarti, T., Madlazim, M., Wasis, W., & Rizki, I. A. (2024). Research trend of Rasch model in the development of scientific literacy assessment through bibliometrics. *Jurnal IPA dan Pembelajaran IPA, 8*(4), 413–429. https://doi.org/10.24815/jipi.v8i4.42307

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice, 17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tugirin, T., Masak, M., Husen, M., & Kuswanto, H. (2025). A systematic literature review on critical thinking skills in physics education. *International Journal of Education in Mathematics, Science and Technology, 13*(5), 1268–1282. https://doi.org/10.46328/ijemst.5709

Utama, R., & Toifur, M. (2025). Investigating senior high school students' conceptual understanding of heat and temperature: A qualitative analysis in physics education. *Jurnal Pendidikan Fisika, 13*(2), 244–258. https://doi.org/10.24127/jpf.v13i2.14137

Zivanka, S. R., Silitonga, T. H., & Silitonga, M. (2022). The application of the Rasch model to develop a two-tier multiple-choice test to measure higher-order thinking skills on motion and force. *Journal of Innovative Science Education, 11*(1), 36–47. https://doi.org/10.21009/JISAE.111.05