JPF | Volume 14 | Number 2 | 2026 | 351 - 371

p - ISSN: 2302-8939
e - ISSN: 2527-4015

Jurnal Pendidikan Fisika

https://journal.unismuh.ac.id/index.php/jpf

DOI: 10.26618/063d8z95

Analysis of Cognitive Ability Instruments Test Based on Marzano's Taxonomy on Temperature and Heat Material Using the Rasch Model

Hani Nur Azizah*, Muslim, Duden Saepuzaman, Lina Aviyanti

Master Program of Physics Education, Universitas Pendidikan Indonesia, Bandung, 40154, Indonesia

**Corresponding author: haninurazzh.25@upi.edu*

Received: January 06, 2026; Accepted: April 09, 2026; Published: May 04, 2026

Abstract – *This study addresses the need for a valid and reliable instrument to assess students' cognitive abilities in physics, particularly in the areas of temperature and heat, which are often associated with conceptual difficulties and misconceptions. The study aimed to evaluate the quality of a cognitive ability test instrument developed based on Marzano's taxonomy by applying the Rasch measurement model. A quantitative design was employed, involving 138 high school students in grades XI and XII from one school in Bandung City, comprising 74 females and 64 males. The instrument comprised 25 multiple-choice items representing five cognitive aspects of Marzano's taxonomy: retrieval, comprehension, analysis, knowledge utilization, and metacognition. Data were analyzed using Winsteps 3.73 to examine item fit, item difficulty, unidimensionality, reliability, person–item distribution, and Differential Item Functioning (DIF) based on gender. The results showed that the instrument generally met Rasch model expectations, with good internal consistency (Cronbach's alpha = 0.76), very high item reliability (0.92), and fair person reliability (0.69). Most items fit the model, although several items showed overfit or unexpected response patterns and require refinement. The item difficulty distribution was dominated by difficult items; the raw variance explained by the measures was 20.3%, and the Wright map indicated that the instrument was reasonably aligned with students' ability levels, though it was less optimal for very high-ability students. DIF analysis showed that most items were gender-neutral, while a small number indicated potential differential functioning. The novelty of this study lies in the systematic operationalization of Marzano's taxonomy into item construction and its evaluation using Rasch analysis in the context of temperature and heat. Overall, the instrument is sufficiently valid for measuring cognitive ability and provides a useful contribution to physics education by offering a psychometrically informed framework for developing more rigorous and meaningful assessment instruments.*

Keywords: *cognitive ability; heat temperature; Marzano taxonomy; physics education; Rasch model.*

I. INTRODUCTION

Science education, particularly physics, plays an important role in fostering students' critical thinking and problem-solving skills. To evaluate the development of these skills, valid

and reliable measurement instruments are needed to accurately map students' cognitive abilities. Cognitive ability refers to an individual's thinking processes in connecting, evaluating, and interpreting events (Mardatila et al., 2020). Within the curriculum context, cognitive aspects constitute core learning objectives; therefore, educators need to design learning experiences that effectively facilitate their development (Dahlan et al., 2021). Cognitive abilities are essential for enhancing students' thinking skills, and quality education can be achieved by engaging students across all levels of the cognitive domain in each lesson (Nabilah et al., 2020). Hardianti (2018) emphasized that analyzing students' cognitive abilities is important for determining the extent to which learning outcomes have been achieved and for identifying students' levels of cognitive attainment. Such analysis can help teachers recognize students' cognitive levels, improve their problem-solving patterns, and support the optimal development of these abilities. One common way to measure students' cognitive abilities is through testing, which provides information needed to evaluate and improve the learning process (Nabilah et al., 2020).

However, evaluation methods reveal that many tools used to assess students' conceptual understanding are not yet fully reliable or valid, as noted by Ceran and Ates (2020) and Irawan et al. (2025). These instruments often fail to capture the full range of students' thinking processes, which can vary widely among individuals. Because of this, the results obtained from such assessments may not accurately represent students' true cognitive abilities or understanding. Consequently, educators might not get a clear or complete picture of students' learning progress, leading to potential misjudgments about their knowledge and skills.

One framework for measuring cognitive abilities more comprehensively is Marzano's taxonomy. This taxonomy not only classifies cognitive processes but also incorporates metacognitive and affective systems. Marzano groups thinking processes into six indicators: retrieval, comprehension, analysis, knowledge utilization, metacognition, and self-system thinking, as well as four types of knowledge: information, mental procedures, physical procedures, and complex skills (Marzano & Kendall, 2007). Compared with Bloom's hierarchical taxonomy, Marzano's taxonomy is considered more flexible and better at representing the dynamics of students' thinking processes (Hattie & Donoghue, 2016; Rawat et al., 2023). It is also more comprehensive because it takes into account how students motivate themselves, regulate their thinking processes, and apply learning strategies (Marzano & Kendall, 2007), thereby helping teachers design learning objectives and activities that are more relevant to students' needs (Cervantes-Pérez et al., 2021; Titova et al., 2023).

Nevertheless, previous studies have generally used Marzano's taxonomy only at a conceptual level and have not systematically operationalized it in the development of test instruments. Furthermore, the relationship between the indicators in Marzano's taxonomy and

item construction has often not been explicitly described. Consequently, the resulting instruments may not fully represent the cognitive dimensions intended to be measured (Cervantes-Pérez et al., 2021; Titova et al., 2023).

In this study, each indicator in Marzano's taxonomy was operationalized into test items. For example, the comprehension indicator was represented by items requiring students to interpret concepts of temperature and heat, whereas the analysis indicator was represented by items requiring students to compare or relate various thermal phenomena. In this way, the instrument construction was not only theoretically grounded but also systematically structured at the item level. To evaluate instrument quality, Classical Test Theory (CTT) is often used to estimate parameters and reliability from observed scores and their variance. However, CTT has several limitations (Ayoola & Ibrahim, 2024). First, student characteristics and item characteristics cannot be separated. Second, the level of item difficulty and overall test difficulty depends heavily on the sample of respondents; as a result, more difficult tests tend to produce lower average scores, whereas easier tests tend to produce higher average scores. Third, tests developed for one group at a given ability level cannot be applied directly to other groups at different ability levels (Ayoola & Ibrahim, 2024).

Therefore, to obtain a more objective and accurate representation of individual student abilities, a quantitative approach using the Rasch model is highly relevant. The Rasch model enables a more detailed analysis of both student abilities and item characteristics. As a modern measurement approach, it can overcome many of the limitations of classical theory. The Rasch model examines multiple aspects, including response patterns, item and respondent fit, dimensionality, item difficulty, and item bias (Irawan et al., 2025). This approach provides information not only about item difficulty but also about each student's ability on the same logit scale (Bond & Fox, 2015). Thus, the Rasch model can identify how accurately items measure cognitive ability and detect response patterns such as guessing or misconceptions.

The Rasch model offers a more objective measurement approach than Classical Test Theory because it produces sample-independent parameter estimates, in which item parameters are independent of participants' abilities (Boone & Staver, 2020). In addition, it supports robust unidimensionality analysis and provides detailed information on item and person fit, enabling evaluation of the consistency between empirical data and the model (Birnbaum et al., 2021). The model also allows item difficulty and respondent ability to be mapped onto a common logit scale, facilitating meaningful interpretation of competence levels (Ismail et al., 2021). In educational assessment contexts, the Rasch model contributes to the development of valid, reliable, and adaptive instruments by providing interval-level data that allow more accurate comparisons across individuals and groups (Dorans & Cook, 2016). Therefore, this approach is particularly

appropriate for analyzing students' cognitive abilities and identifying aspects of the instrument that require further improvement.

The topic of temperature and heat is part of the high school physics curriculum and often requires students not only to understand concepts factually but also to relate them to everyday phenomena through higher-order thinking (Abraham et al., 2021). Temperature and heat are also among the physics concepts that frequently give rise to misconceptions and are closely related to students' daily experiences (Haryono & Aini, 2021). Several studies have shown that students tend to perceive temperature and heat as abstract concepts, which can lead to varied interpretations and persistent misunderstandings when they encounter the material again (Baser, 2006; Aprilia & Dwandaru, 2024). Therefore, an evaluative approach is needed that can reveal, in depth, the level of students' cognitive abilities in understanding physics concepts, particularly those related to temperature and heat.

Based on this background, this study aims to analyze a cognitive ability test instrument grounded in Marzano's taxonomy of temperature and heat, using the Rasch model as the analytical tool. To guide the study, the following research questions were formulated: (1) How are the validity and reliability of the cognitive ability test instrument based on Marzano's taxonomy? (2) How are students' ability distribution and the difficulty classification of the cognitive ability test items based on Marzano's taxonomy? (3) Is there any Differential Item Functioning (DIF) based on gender?

II. METHODS

This study employed a quantitative research design to analyze the quality of a cognitive ability test instrument developed on the topic of temperature and heat. Quantitative research is appropriate for studies that systematically examine data using structured instruments and measurable procedures, particularly when the objective is to empirically evaluate variables and produce scientifically accountable findings (Creswell & Creswell, 2018). In this study, the Rasch model was used as the main analytical framework because it provides a more comprehensive evaluation of item functioning and respondent ability than conventional scoring approaches. Rasch analysis is especially useful for identifying item fit, person fit, item difficulty, dimensionality, and possible item bias, thereby strengthening the psychometric quality of educational instruments (Boone & Staver, 2020).

The participants in this study were selected through purposive sampling, a technique that emphasizes the intentional selection of individuals or sites based on characteristics relevant to the research purpose (Creswell & Creswell, 2018). The study involved 138 students in grades XI and

XII from a senior high school in Bandung City who had previously studied temperature and heat. Of the total participants, 74 were female and 64 were male, with an average age range of 16 to 17 years. The use of a single school context allowed the researchers to maintain relative consistency in curriculum exposure and instructional background, thereby reducing external variability in students' responses. Although this sampling frame limits the broader generalizability of the findings, the number of respondents was considered sufficient for Rasch modeling. Mešić et al. (2019) suggested that at least 100 participants are needed to ensure stable parameter estimation in Rasch-based concept understanding tests; therefore, the sample size in this study met the minimum requirement for robust analysis.

The research instrument was a multiple-choice cognitive ability test designed to measure students' understanding of temperature and heat based on Marzano's taxonomy. The instrument consisted of 25 multiple-choice items, each with five answer options, and was developed to represent five cognitive aspects: retrieval, comprehension, analysis, knowledge utilization, and metacognition (Marzano & Kendall, 2007). Instrument development was conducted through several systematic stages. The first stage was construct formulation, in which indicators and sub-indicators of cognitive ability were identified based on Marzano's taxonomy and aligned with the content of temperature and heat. The second stage was item development, during which the researchers constructed multiple-choice items according to the predetermined blueprint. Each item was written to reflect the intended cognitive process and content focus. The third stage involved content validation by two physics education lecturers. This expert review was intended to evaluate the items' relevance to the targeted indicators, their appropriateness for high school physics content, and their clarity in language, construction, and spelling. Revisions were then made in response to expert feedback until the instrument met the expected content validity criteria. The fourth stage was a limited pilot test, conducted with one science class of 30 students whose characteristics were similar to those of the main sample. This pilot study was conducted to assess the instrument's empirical quality prior to the main administration.

Each item was scored dichotomously, with 1 assigned to correct answers and 0 to incorrect ones. The distribution of questions developed is summarized in the following Table 1

Table 1. Distribution of questions based on Marzano's taxonomy

Level	Indicators	Information	Item
Retrieval	Recognizing	Students recognize features of information but do not necessarily understand the structure of knowledge or differentiate between critical and noncritical components.	24
	Recalling	Students produce features of information but do not necessarily understand the structure of knowledge or differentiate between critical and noncritical components.	1,6

Comprehension	Symbolizing	Students construct an accurate symbolic representation of the knowledge, differentiating critical and noncritical components	4,5,22,25
	Integrating	Students identify the basic structure of knowledge and the critical as opposed to noncritical characteristics.	15
Analysis	Specifying	Students identify specific applications or logical consequences of the knowledge.	2,3,9,14
	Generalizing	Students construct new generalizations or principles based on the knowledge.	23
	Analyzing errors	Students identify errors in the presentation or use of the knowledge.	19,20
	Classifying	Students identify superordinate and subordinate categories related to the knowledge.	17
	Matching	Students identify important similarities and differences between knowledge components.	18
Utilization knowledge	Investigating	Students use the knowledge to conduct investigations or to investigate the knowledge.	7,8
	Experimenting	Students use the knowledge to generate and test hypotheses or generate and test hypotheses about the knowledge	10
	Problem solving	Students use the knowledge to solve problems or to solve problems about the knowledge.	11,13
Metacognition	Decision making	Students use the knowledge to make decisions, or to make decisions about the knowledge.	12
	Monitoring accuracy	Students determine the extent to which they are accurate about their knowledge	24
	Monitoring clarity	Students determine the extent to which they have clarity about the knowledge.	21

The following is an example of a question item (S4) developed in the comprehension aspect, with the sub-material on expansion.

Sub Material : Expansion

Cognitive Aspect : Comprehension

Question Indicator : Students can construct accurate symbolic representations related to the physical equations of objects experiencing volume expansion.

Question Items :

In the design process of a precision machine, a metal cube is used as part of the internal support frame. This cube has a volume x and is made of a metal material with a coefficient of linear expansion of y. The machine will operate at a significantly higher temperature. One thing that needs to be considered is the increase in the surface area of the cube due to an increase in temperature of z when the machine is operating. If the temperature of the cube is increased by z, then the surface area of the cube will increase by ….

A. xyz
B. $6yxz$
C. $12zyx$
D. $6y\,x^{2/3}\,z$
E. $12y\,x^{2/3}z$

Figure 1. Examples of developed test items

The data obtained from the main administration were analyzed using Winsteps version 3.73. Rasch analysis was performed to examine item validity, instrument reliability, item

difficulty, respondent ability distribution, dimensionality, and Differential Item Functioning (DIF). Item validity was evaluated through item fit statistics, specifically Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (PT Measure Corr). This analysis aims to ensure that each item functions in accordance with the Rasch model, thereby validly measuring the intended construct. The item acceptance criteria are based on the values presented in Table 2.

Table 2. The criteria for the value of each question item

Indicator	Criteria
Outfit MNSQ	0,5 < MNSQ < 1,5
Outfit ZSTD	-2 < ZSTD < +2
PT Measure Corr	0,4 < PT Measure Corr < 0,85

(Sumintono & Widhiarso, 2015)

Next, the scores for each student were analyzed using a unidimensionality test. The unidimensionality test was conducted to ensure that the instrument measures a single construct, namely, cognitive ability. This was analyzed through the Raw Variance Explained by Measures value in the dimensionality output. This test is important because the Rasch Model requires the assumption of unidimensionality for the measurement results to be construct valid. Interpretation of the results is based on the criteria in Table 3.

Table 3. Interpretation of the unidimensionality of the instrument

Raw variance explained by measures	Interpretation
> 20%	Fulfilled
>40%	Appropriate
>60%	Excellent

(Sumintono & Widhiarso, 2015)

The parameters measured in this study include construct identification carried out to ensure that each question item represents an aspect that is in accordance with the concept being measured (Boone et al., 2014). The reliability and consistency of the instrument were analyzed using the Rasch model, including reliability coefficients and respondent suitability (fit person) statistics. The reliability criteria for test instruments in the Rasch Model, according to Sumintono and Widhiarso (2015), are shown in Table 4 as follows:

Table 4. Interpretation of reliability value criteria

Criteria (person/item)	Interpretation
r > 0.94	Excellent
$0.91 \leq r \leq 0.94$	Very good
$0.81 \leq r \leq 0.90$	Good
$0.67 \leq r \leq 0.80$	Fair
r < 0.67	Poor

(Sumintono & Widhiarso, 2015)

Then, there is Cronbach's Alpha, which is one of the indicators to measure the internal reliability of an instrument, which shows the extent to which items in a test measure the same construct consistently (Tavakol & Dennick, 2011). The Cronbach's Alpha criteria in the Rasch Model, according to Sumintono and Widhiarso (2015), are shown in Table 5 as follows:

Table 5. Interpretation of Cronbach's alpha values

Cronbach's alpha	Interpretation
$\alpha \geq 0.80$	Excellent
$0.70 \leq \alpha < 0.80$	Good
$0.60 \leq \alpha < 0.70$	Acceptable
$0.50 \leq \alpha < 0.60$	Fair
$\alpha < 0.50$	Poor

(Sumintono & Widhiarso, 2015)

In addition, this study also analyzed the Wright Map, a measure of student ability (Person Measure). The Wright Map is used to visualize the distribution of student ability and item difficulty on a single logit scale. This analysis assesses whether the item difficulty level matches the respondent's ability, enabling a more in-depth evaluation of the instrument's quality. Person Fit analysis was conducted to identify respondents whose response patterns were inconsistent with the Rasch model. This is crucial for ensuring data quality and detecting potential invalid responses (Syamsyiah et al., 2023), while the Person Measure was used to assess students' abilities. Respondent ability was estimated from item response patterns, yielding objective, comparable logit scores (Bond & Fox, 2015). This approach allows for a more objective comparison of students' abilities than raw scores do.

III. RESULTS

This study used the Rasch model to evaluate the quality of a cognitive ability test instrument based on Marzano's Taxonomy on the topic of temperature and heat. The analysis focused on item fit, item difficulty, unidimensionality, reliability, the Wright map, and Differential Item Functioning (DIF) based on gender. Overall, the results provide evidence on the instrument's psychometric quality and identify several aspects that require further refinement.

Item fit analysis

The first stage of the analysis examined the fit of each item to the Rasch model. Item fit analysis was conducted to determine whether the items functioned appropriately in measuring the intended cognitive ability construct. In Rasch measurement, fit statistics indicate the extent to which observed responses correspond to model expectations. In this study, item fit was evaluated using MNSQ, ZSTD, and PT Measure Corr.

The total score is the sum of the raw scores of all respondents who answered the item. Total count is the number of respondents who answered the item. Size refers to the Rasch item size in logit units. Lower and higher item sizes represent more and less interesting items, respectively. Model SE refers to the standard error of the item size in logit units. Outfit MNSQ refers to a fit statistic that is sensitive to extreme responses. Infit MNSQ refers to a fit statistic that uses a weighted average (Zoechling et al., 2022).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.		EXACT MATCH EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
24	106	138	.09	.22	1.25	2.1	1.47	2.3	A	.11	.37	75.4	78.9	S24
23	100	138	.36	.21	1.15	1.5	1.31	1.8	B	.22	.38	70.3	75.7	S23
19	115	138	-.39	.25	1.16	1.1	1.23	1.0	C	.19	.34	82.6	84.2	S19
17	93	138	.65	.20	1.12	1.4	1.17	1.3	D	.28	.39	66.7	72.4	S17
10	114	138	-.33	.24	1.16	1.1	.95	-.1	E	.25	.35	79.0	83.6	S10
22	112	138	-.22	.23	1.14	1.1	1.00	.1	F	.25	.35	78.3	82.4	S22
20	87	138	.88	.19	1.06	.9	1.14	1.2	G	.33	.40	69.6	70.0	S20
21	106	138	.09	.22	1.12	1.0	1.05	.4	H	.27	.37	75.4	78.9	S21
18	96	138	.53	.20	1.04	.5	1.10	.7	I	.34	.39	73.2	73.7	S18
7	106	138	.09	.22	1.06	.6	1.05	.3	J	.31	.37	79.7	78.9	S7
25	96	138	.53	.20	1.01	.1	.96	-.2	K	.39	.39	74.6	73.7	S25
3	95	138	.57	.20	.98	-.2	.95	-.3	L	.41	.39	73.9	73.2	S3
6	84	138	.99	.19	.96	-.6	.97	-.2	M	.44	.40	72.5	69.0	S6
13	108	138	-.01	.22	.97	-.2	.91	-.4	l	.40	.36	79.7	80.0	S13
12	79	138	1.16	.19	.96	-.6	.93	-.7	k	.44	.40	68.1	67.7	S12
4	130	138	-1.71	.38	.95	-.1	.56	-.8	j	.33	.24	94.2	94.2	S4
5	70	138	1.48	.19	.94	-1.0	.93	-.7	i	.46	.40	68.8	66.8	S5
15	105	138	.14	.22	.93	-.6	.88	-.6	h	.43	.37	80.4	78.3	S15
9	113	138	-.27	.24	.91	-.6	.85	-.6	g	.43	.35	84.1	83.0	S9
16	90	138	.76	.20	.89	-1.4	.82	-1.5	f	.51	.39	74.6	71.1	S16
1	95	138	.57	.20	.87	-1.5	.83	-1.2	e	.51	.39	79.7	73.2	S1
14	115	138	-.39	.25	.86	-.9	.64	-1.6	d	.49	.34	85.5	84.2	S14
11	133	138	-2.24	.47	.82	-.4	.37	-1.0	c	.38	.20	96.4	96.4	S11
2	131	138	-1.86	.40	.77	-.6	.30	-1.6	b	.46	.23	94.9	94.9	S2
8	128	138	-1.45	.34	.75	-.9	.41	-1.6	a	.50	.26	93.5	92.8	S8
MEAN	104.3	138.0	.00	.24	.99	.1	.91	-.2				78.8	79.1	
S.D.	16.0	.0	.93	.07	.13	1.0	.28	1.0				8.5	8.4	

Figure 2. Item Fit Analysis

Based on Figure 2**,** the average outfit value was 0.91, indicating that, in general, the instrument met the acceptable fit criteria. Most items fit the Rasch model consistently and were therefore considered appropriate for measuring the intended construct. However, several items showed indications of misfit or overfit. Items S11, S2, and S8 had Outfit MNSQ values below the lower threshold, namely 0.37, 0.30, and 0.41, respectively. These results indicate overfit, suggesting that the response patterns for these items were overly predictable and contributed less information to the measurement process. In addition, item S24 showed a high Outfit ZSTD value of 2.3, indicating an unusual or less predictable response pattern. This result suggests that the item may be ambiguous, contextually inconsistent, or worded in a way that led students to respond unexpectedly. Overall, the item fit analysis indicates that the instrument functioned adequately, although several items should be reviewed more carefully to improve measurement quality.

Item difficulty analysis

Item difficulty was analyzed using the Rasch measure value expressed in logits. The classification of item difficulty is summarized in Table 6. Analysis of the results of the level of difficulty of the questions**.** The analysis was based on the standard deviation of the item measure distribution (SD = 0.93), with items classified as very easy, easy, difficult, or very difficult according to their logit positions.

Table 6. Analysis of the results of the level of difficulty of the questions

Measure logit (SD = 0,93)	Interpretation of the difficulty level	Item	Number of items
M < -SD	Very easy	S2, S4, S8, S11	4
-SD ≤ M ≤ 0	Easy	S9, S10, S13, S14, S21	5
0 ≤ M ≤ SD	Difficult	S1, S3, S7, S15, S16, S17, S18,19, S22, S23, S24, S25	12
M > SD	Very difficult	S6, S20, S5, S12	4

(Adams et al., 2022)

As shown in Table 6, four items were classified as very easy, namely S2, S4, S8, and S11, representing 16% of the total items. Five items were categorized as easy: S9, S10, S13, S14, and S21, which accounted for 20% of the instrument. The largest proportion of items fell into the difficult category, comprising 12 items (48% of the total): S1, S3, S7, S15, S16, S17, S18, S19, S22, S23, S24, and S25. The remaining four items, namely S6, S20, S5, and S12, were classified as very difficult, representing 16% of the instrument. These findings indicate that the distribution of item difficulty was not fully balanced, as the instrument was dominated by items in the difficult category. This pattern suggests that the test tended to emphasize more demanding cognitive processes, although the spread of items across difficulty levels remained sufficient to capture variation in student performance.

Unidimensionality

The instrument's dimensionality was examined to determine whether it predominantly measured a single underlying construct. This analysis was based on the standardized residual variance output, as shown in Figure 3. Standardized residual variance. In Rasch measurement, unidimensionality is an important requirement because it indicates that the instrument primarily assesses a single dominant trait.

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --		Modeled
Total raw variance in observations	=	31.4	100.0%	100.0%
Raw variance explained by measures	=	6.4	20.3%	21.1%
Raw variance explained by persons	=	3.2	10.2%	10.6%
Raw Variance explained by items	=	3.2	10.1%	10.5%
Raw unexplained variance (total)	=	25.0	79.7% 100.0%	78.9%
Unexplned variance in 1st contrast	=	2.3	7.4% 9.3%	
Unexplned variance in 2nd contrast	=	1.7	5.5% 7.0%	
Unexplned variance in 3rd contrast	=	1.6	5.1% 6.5%	
Unexplned variance in 4th contrast	=	1.5	4.9% 6.2%	
Unexplned variance in 5th contrast	=	1.5	4.7% 5.9%	

Figure 3. Standardized residual variance

Based on Figure 3, the raw variance explained by measures was 6.4, equivalent to 20.3%. This result indicates that the variance explained by the Rasch model met the minimum threshold for unidimensionality, although the measured construct's contribution to the total variance remained relatively limited. In other words, the instrument showed evidence of measuring a primary construct, but student responses may also have been influenced by additional factors outside the intended dimension. Therefore, the instrument can be considered to have achieved a minimum acceptable level of unidimensionality, while still leaving room for improvement in order to strengthen construct coherence.

Reliability

The reliability analysis in the Rasch model included person reliability, item reliability, and Cronbach's alpha. These results are presented in Figure 4. Reliability and Cronbach's Alpha. Together, these indicators provide a more comprehensive picture of the instrument's consistency and stability.

SUMMARY OF 138 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	18.9	25.0	1.50	.56	1.01	.1	.91	.0
S.D.	3.8	.0	.98	.15	.11	.5	.23	.5
MAX.	24.0	25.0	3.49	1.03	1.46	1.4	1.71	2.2
MIN.	5.0	25.0	-1.62	.43	.75	-1.7	.34	-1.2
REAL RMSE	.59	TRUE SD	.79	SEPARATION	1.33	Person RELIABILITY		.69
MODEL RMSE	.58	TRUE SD	.79	SEPARATION	1.37	Person RELIABILITY		.70
S.E. OF Person MEAN = .08								

Person RAW SCORE-TO-MEASURE CORRELATION = .97

CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .76

SUMMARY OF 25 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	104.3	138.0	.00	.24	.99	.1	.91	-.2
S.D.	16.0	.0	.93	.07	.13	1.0	.28	1.0
MAX.	133.0	138.0	1.48	.47	1.25	2.1	1.47	2.3
MIN.	70.0	138.0	-2.24	.19	.75	-1.5	.30	-1.6
REAL RMSE	.26	TRUE SD	.90	SEPARATION	3.47	Item RELIABILITY		.92
MODEL RMSE	.25	TRUE SD	.90	SEPARATION	3.54	Item RELIABILITY		.93
S.E. OF Item MEAN = .19								

Figure 4. Reliability and Cronbach's Alpha

As shown in Figure 4, the Cronbach's alpha value was 0.76, indicating good internal consistency. This suggests that the items were sufficiently consistent in measuring the same general construct. The item reliability value was 0.92, which falls into the very good category and indicates that the hierarchy of item difficulty was stable and well defined. By contrast, the person reliability value was 0.69, which is categorized as fair. This result shows that the instrument had a moderate ability to distinguish among students with different levels of cognitive ability. Taken together, these findings indicate that the instrument performed well in terms of internal consistency and item stability, although its capacity to differentiate student ability levels could still be improved.

Wright map

The Wright map was used to display the distribution of student ability and item difficulty on the same logit scale. This analysis is presented in Figure 5. The Wright map enables direct comparison between the level of respondent ability and the level of item difficulty. On the map, students with higher cognitive ability are positioned in the upper-left section, while lower-ability

students appear in the lower-left section. On the right side, more difficult items are located higher on the scale, whereas easier items are located lower.

Wright's map displays the distribution of item difficulty and student ability on a logit scale (Bond & Fox, 2015).

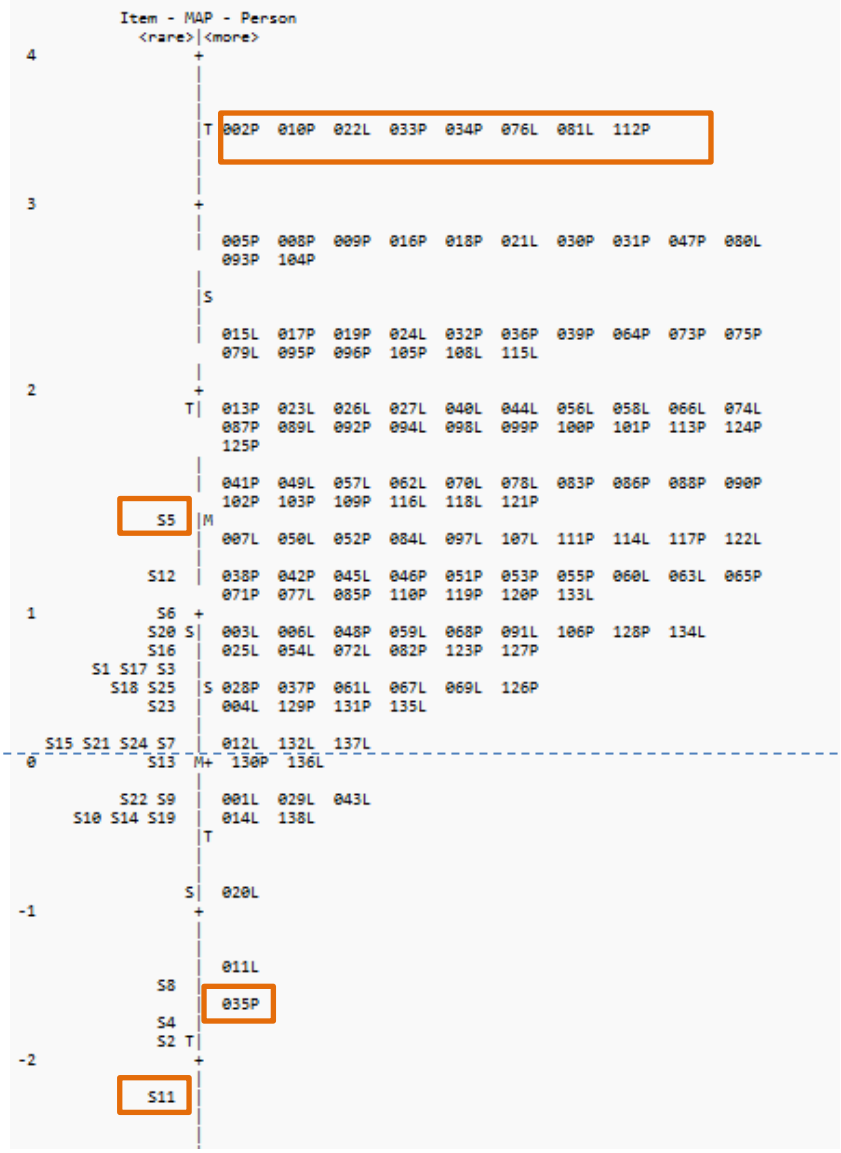

Figure 5. Wright map

Based on Figure 5, individual ability estimates ranged from approximately -2 to +4 logits, with most students clustered between +1 and +3 logits. Item difficulty estimates ranged from approximately -2 to +2 logits, with the most difficult items located near +2 logits and the easiest items near -2 logits. This distribution indicates that the instrument included items spanning a relatively broad range of difficulty levels and was generally able to differentiate students across different ability levels on the topic of temperature and heat. At the same time, the map also

suggests that a number of students with relatively high ability were located above the difficulty range of most items, indicating that the instrument may provide limited discrimination for the highest-performing respondents. Nevertheless, the Wright map overall shows a reasonably good alignment between item difficulty and student ability.

Differential item functioning (DIF)

Differential Item Functioning analysis was conducted to examine whether certain items functioned differently for male and female students. In Rasch analysis, DIF is used to detect potential item bias across groups that may not be attributable to differences in the construct being measured. The results of this analysis are shown in Figure 7. Person (gender) DIF analysis.

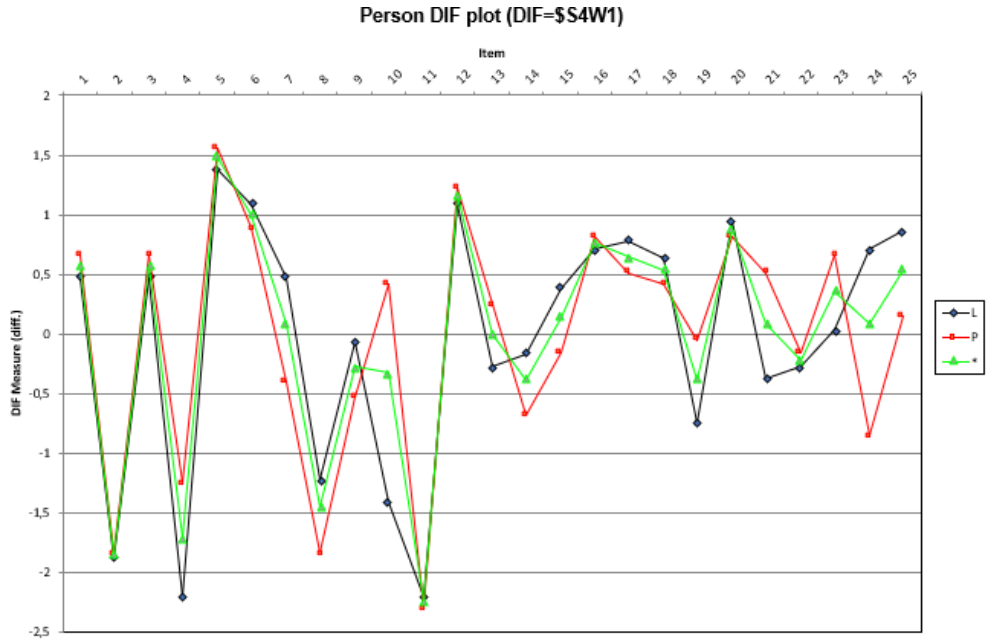

Figure 7. Person (gender) DIF analysis

Based on Figure 7, most items had DIF values between -0.5 and +0.5 logits, indicating they were relatively neutral across gender groups. This result suggests that, overall, male and female students had similar probabilities of responding correctly to most items when their underlying ability was taken into account. However, several items showed moderate-to-high DIF. Items S14, S15, and S24 had positive DIF values, indicating that these items were relatively more difficult for female students. In contrast, items S10, S13, S19, and S21 had negative DIF values, indicating that they were relatively easier for female students than for male students. These findings show that although the instrument was generally fair, several items may exhibit differential functioning by gender and therefore warrant further review.

IV. DISCUSSION

The findings of this study provide empirical evidence regarding the quality of a cognitive ability test instrument developed based on Marzano's Taxonomy and analyzed using the Rasch model. Overall, the results indicate that the instrument has an adequate psychometric foundation, although several aspects still require refinement to improve measurement precision, construct representation, and fairness. These findings are important because the quality of an assessment instrument depends not only on its theoretical framework but also on the extent to which its items function consistently and meaningfully in actual measurement contexts.

The item fit analysis showed that most items met the Rasch model criteria, suggesting that the instrument generally functioned well in measuring the intended construct. This result indicates that, at the structural level, the instrument has acceptable internal validity. In Rasch measurement, fit statistics are essential because they indicate whether observed response patterns align with model expectations. Most items were found to be suitable for measurement, supporting the appropriateness of operationalizing Marzano's taxonomy as test items on the topic of temperature and heat. However, several items, namely S2, S8, and S11, showed Outfit MNSQ values below 0.5, indicating overfit. In Rasch analysis, overfitting does not necessarily represent a serious flaw, but it suggests that the responses to these items were too predictable and therefore contributed limited additional information to distinguishing among students (Linacre, 2006).

This pattern may be explained by the fact that some items measured highly familiar or straightforward concepts. When an item is overly explicit or cognitively undemanding, students with different ability levels may respond in very similar ways, thereby reducing the item's discriminatory value. The original analysis of item S8 illustrates this condition well. The item addressed the basic concept of thermal equilibrium, in which heat moves from a higher-temperature object to a lower-temperature object until thermal balance is reached. Because this concept is fundamental and often well understood by students, many respondents answered correctly with little difficulty. As a result, the item became highly predictable and less informative for distinguishing subtle differences in cognitive ability. This interpretation is consistent with the Rasch view that items should not only be content-correct but also sufficiently informative across a range of respondent abilities (Amelia, 2021).

A different issue emerged in item S24, which displayed an Outfit ZSTD value outside the acceptable range. This result suggests an unusual response pattern, indicating that students' answers to this item were less predictable than the model expected. Such a misfit may stem from item ambiguity, excessive contextual complexity, or competing reasoning strategies among students. In this study, item S24 required students to interpret a heat-transfer experiment and relate

it to the thermal conductivity of different metals. This cognitive demand is conceptually appropriate for higher-level assessment, but it may also introduce interpretive difficulty if the wording or context is not sufficiently focused. Some students may have relied on everyday intuition rather than formal conceptual reasoning, while others may have attended selectively to certain variables presented in the problem. Consequently, the variability in response patterns may reflect not only conceptual understanding but also differences in how students interpreted the item. This finding suggests that revision is needed not to lower the item's cognitive level, but to improve clarity so that the item more accurately captures the intended construct.

The analysis of item difficulty showed that the instrument was dominated by items in the difficult category. This pattern is consistent with the intention to develop an instrument based on Marzano's taxonomy, which emphasizes not only retrieval and comprehension but also analysis, knowledge utilization, and metacognition. In this sense, the distribution of difficulty levels may be interpreted as evidence that the instrument attempts to move beyond low-level recall toward higher-order thinking. However, the distribution was not fully balanced. The relatively small number of very easy and very difficult items indicates that the instrument does not yet represent the full spectrum of student ability equally well. This imbalance has methodological consequences, as an instrument with limited coverage at the extremes may be less effective at identifying students with very low or very high ability. Thus, while the existing item set is suitable for measuring students in the middle-to-upper ability range, additional refinement is needed to improve targeting across the full continuum of cognitive performance.

The unidimensionality results further support a cautious but positive interpretation of the instrument's construct validity. The raw variance explained by measures was 20.3%, which meets the minimum threshold suggested by Sumintono and Widhiarso (2015), but still indicates that the model explains only a limited portion of the total variance. This result suggests that the instrument is oriented toward measuring one dominant construct, yet the responses may still be influenced by other dimensions. Such a finding is understandable in the context of Marzano's taxonomy, because the framework itself encompasses several distinct but related thinking processes. When an instrument integrates retrieval, comprehension, analysis, knowledge utilization, and metacognition into a single test, some degree of heterogeneity in response processes may naturally emerge. Therefore, the present result does not invalidate the instrument, but it does indicate that stronger alignment across items may be needed if greater construct coherence is desired.

The reliability findings provide additional insight into the instrument's quality. The Cronbach's alpha value of 0.76 indicates good internal consistency, meaning that the items were sufficiently related in measuring the same overall construct. The item reliability value of 0.92 further indicates that the hierarchy of item difficulty was stable and replicable across the sample.

This is an important strength, as it suggests that the relative positioning of items on the logit scale is dependable. However, the person's reliability value of 0.69 was only in the fair category. This indicates that the instrument's ability to distinguish among different levels of student ability was adequate but not strong. One likely reason for this pattern is the uneven distribution of item difficulty, which may have limited the variability of student responses. When items are clustered within a relatively narrow difficulty band or do not sufficiently target the extremes of ability, the instrument becomes less effective in classifying respondents into more refined ability levels. Therefore, although the instrument already demonstrates solid internal consistency and stable item functioning, its ability to discriminate among students remains an area for further development.

The Wright map analysis provides a particularly useful perspective on the relationship between student ability and item difficulty. The map showed that most items were distributed along the logit scale, broadly overlapping with the respondents' ability distribution. This suggests that, in general, the instrument was reasonably well targeted to the sample. At the same time, the Wright map revealed that some high-ability students were positioned above the range of most test items. This indicates that the instrument was less effective in differentiating among the highest-performing students, because there were not enough sufficiently difficult items to challenge them. This result is closely related to the person reliability finding and reinforces the interpretation that future revisions should include more items targeting advanced cognitive processes. In particular, item development at the levels of knowledge utilization and metacognition could be expanded to better represent complex reasoning and provide stronger measurement at the upper end of the ability scale.

The Differential Item Functioning analysis adds an important dimension to the interpretation of instrument quality by addressing fairness. Most items showed DIF values between -0.5 and +0.5 logits, indicating that they functioned similarly for male and female students. This result supports the instrument's general fairness and suggests that most items measured cognitive ability rather than group membership. However, several items showed moderate-to-high DIF, indicating potential bias. Items S14, S15, and S24 appeared relatively more difficult for female students, while items S10, S13, S19, and S21 appeared relatively easier for female students than for male students. These patterns do not necessarily imply unfairness in an absolute sense, but they do indicate that certain item characteristics may interact differently with students across gender groups. Such interaction may result from contextual familiarity, wording, representation format, or reasoning demands unrelated to the target construct. In line with Bond and Fox (2015), items showing substantial DIF should be reviewed carefully to ensure

that differences in performance reflect cognitive ability rather than extraneous group-related influences.

Taken together, the results of this study indicate that the developed instrument already has sufficient validity and reliability, especially for initial use in assessing cognitive ability related to temperature and heat. At the same time, the findings clearly show that further refinement is necessary. The most important areas for improvement include balancing the distribution of item difficulty, strengthening the construct's coherence across the full set of items, improving the discrimination of respondents' ability levels, and reviewing items that exhibit misfit or gender-related DIF. These refinements are not merely technical adjustments but are essential for ensuring that the instrument is both psychometrically sound and pedagogically meaningful. As such, this study contributes not only to measuring students' cognitive abilities in physics but also to the broader effort to develop assessment instruments that more effectively integrate theoretical frameworks, statistical evidence, and classroom relevance.

V. CONCLUSION AND SUGGESTION

This study examined the quality of a cognitive ability test instrument on the topic of temperature and heat developed based on Marzano's taxonomy and analyzed using the Rasch model. The findings showed that the instrument demonstrated acceptable psychometric quality, as indicated by good internal consistency (Cronbach's alpha = 0.76), very high item reliability (0.92), and acceptable person reliability (0.69). Most items fit the Rasch model, supporting the instrument's validity, although several items showed overfitting or unexpected response patterns and therefore require revision. The analysis of item difficulty indicated that the instrument was dominated by difficult items, suggesting it was more effective at measuring middle- to higher-level cognitive processes than at the extremes of ability. The one-dimensionality result met the minimum acceptable threshold, while the Wright map showed that the instrument generally differentiated students across a range of abilities, though it was less optimal for very high-ability students. In addition, the DIF analysis indicated that most items were gender-neutral, although several showed potential differential functioning that should be further reviewed. Overall, these findings suggest that the instrument is suitable for measuring students' cognitive ability on temperature and heat, while also highlighting the need for targeted refinement to improve precision and fairness.

This study has several limitations that should be considered when interpreting the findings. The participants were drawn from a single secondary school, which limits the generalizability of the results across educational contexts. In addition, the instrument was restricted to multiple-

choice items, which may not fully capture the complexity of higher-order cognitive processes, particularly those related to knowledge utilization and metacognition. Future research is therefore recommended to involve more diverse samples from different schools or regions, develop instruments with a more balanced range of item difficulty levels, and incorporate additional item formats, such as open-ended or performance-based tasks, to capture students' thinking more comprehensively. Further studies are also needed to re-examine items that show indications of misfit or gender-related DIF to strengthen measurement fairness. Despite these limitations, this study contributes to the field of physics education by providing an empirically tested assessment instrument grounded in Marzano's taxonomy and evaluated through Rasch analysis. It also offers a practical, methodological reference for researchers and educators seeking to develop more valid, reliable, and pedagogically meaningful instruments to assess students' cognitive abilities in physics learning.

REFERENCES

Abraham, Y. M., Valentin, M., Hansen, B., Bauman, L. C., & Robertson, A. D. (2021). Exploring student conceptual resources about heat and temperature. *Physics Education Research Conference Proceedings*, 21–26. American Association of Physics Teachers. https://doi.org/10.1119/perc.2021.pr.Abraham

Adams, D., Chuah, K. M., Sumintono, B., & Mohamed, A. (2022). Students' readiness for e-learning during the COVID-19 pandemic in a South-East Asian university: A Rasch analysis. *Asian Education and Development Studies, 11*(2), 324–339. https://doi.org/10.1108/AEDS-05-2020-0100

Amelia, R. N. (2021). Identifikasi item fit dan person fit dalam pengukuran hasil belajar kimia. *Jurnal Ilmiah WUNY, 3*(1), 13–26. https://www.researchgate.net/publication/352097795_identifikasi_item_fit_dan_person_fit_dalam_pengukuran_hasil_belajar_kimia

Aprilia, A., & Dwandaru, W. S. B. D. (2024). Empirical analysis of physics test instruments to measure graphical representation abilities in "temperature and heat" topics. *Science Education International, 35*(3), 240–249. https://www.icaseonline.net/journal/index.php/sei/article/view/923

Ayoola, F. W., & Ibrahim, A. (2024). Item statistics of multiple choice physics achievement test using classic test theory and item response theory. *EAS Journal of Psychology and Behavioural Sciences, 6*(2), 11–18. https://doi.org/10.36349/easjpbs.2024.v06i02.002

Baser, M. (2006). Effect of conceptually change-oriented instruction on students' understanding of heat and temperature concepts. *Journal of Maltese Educational Research, 4*(1), 64–79. https://eric.ed.gov/?id=ED495216

Birnbaum, M., Brock, K., Parkinson, S., Burton, E., Clark, R., & Hill, K. D. (2021). Rasch analysis of the Burke Lateropulsion Scale (BLS). *Topics in Stroke Rehabilitation, 28*(4), 268–275. https://doi.org/10.1080/10749357.2020.1824724

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. (2020). Correction to: Advances in Rasch analyses in the human sciences. *Advances in Rasch analyses in the human sciences*. Springer. https://doi.org/10.1007/978-3-030-43420-5_21

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. http://dx.doi.org/10.1007/978-94-007-6857-4

Ceran, S. A., & Ates, S. (2020). Conceptual understanding levels of students with different cognitive styles: An evaluation in terms of different measurement techniques. *Eurasian Journal of Educational Research, 88*, 149–178. https://www.researchgate.net/publication/343504309_Conceptual_Understanding_Levels_of_Students_with_Different_Cognitive_Styles_An_Evaluation_in_Terms_of_Different_Measurement_Techniques

Cervantes-Pérez, F., Navarro-Perales, J., Franzoni-Velázquez, A. L., & de la Fuente-Valentín, L. (2021). Bayesian knowledge tracing for navigation through Marzano's taxonomy. *International Journal of Interactive Multimedia and Artificial Intelligence, 6*(6), 234–239. https://doi.org/10.9781/ijimai.2021.05.006

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Dahlan, A., Herman, H., & Yani, A. (2021). Analisis kemampuan kognitif dalam menyelesaikan soal-soal fisika peserta didik SMAN 21 Makassar. *Jurnal Sains dan Pendidikan Fisika, 17*(2), 146–152. https://www.neliti.com/publications/485260/analisis-kemampuan-kognitif-dalam-menyelesaikan-soal-soal-fisika-peserta-didik-s

Dorans, N. J., & Cook, L. L. (2016). *Fairness in educational assessment and measurement*. Routledge. https://doi.org/10.4324/9781315774527

Hardianti, T. (2018). Analisis kemampuan peserta didik pada ranah kognitif dalam pembelajaran fisika SMA. *Seminar Nasional Quantum: Jurnal Pendidikan Fisika UAD*, *25,* 557–561. https://www.academia.edu/download/77527040/263.pdf

Haryono, H. E., & Aini, K. N. (2021). Diagnosis misconceptions of junior high school in Lamongan on the heat concept using the three-tier test. *Journal of Physics: Conference Series, 1806*(1), 1-6. https://doi.org/10.1088/1742-6596/1806/1/012002

Hattie, J. A. C., & Donoghue, G. M. (2016). Learning strategies: A synthesis and conceptual model. *npj Science of Learning, 1*, 1-13. https://doi.org/10.1038/npjscilearn.2016.13

Irawan, I. D. A., Indraloka, R. M., Basri, N. A., Salmah, U., & Parno, P. (2025). Analysis of concept understanding test items on static fluid material using Rasch model. *Jurnal Pendidikan Fisika, 13*(1), 1–13. https://doi.org/10.26618/jpf.v13i1.15687

Ismail, M. S., Din, M. S. H., & Jusoh, M. S. (2021). Predictive modelling using Rasch's person-item map: Interpreting and assessing in Malaysia manufacturing firms. *AIP Conference Proceedings, 2347*. American Institute of Physics. https://doi.org/10.1063/5.0052973

Linacre, J. M. (2006). Data variance explained by Rasch measures. *Transactions of the Rasch Measurement SIG, 20*(1). https://www.rasch.org/rmt/rmt201a.htm

Mardatila, A., Novia, H., & Sinaga, P. (2020). Penerapan pembelajaran fisika menggunakan multi-representasi untuk meningkatkan kemampuan kognitif dan pemecahan masalah siswa SMA pada pokok bahasan gerak parabola. *Omega: Jurnal Fisika dan Pendidikan Fisika, 5*(2), 33–39. https://www.researchgate.net/publication/337677739_Penerapan_Pembelajaran_Fisika_Menggunakan_Multi_Representasi_untuk_Meningkatkan_Kemampuan_Kognitif_dan_Pemecahan_Masalah_Siswa_SMA_pada_Pokok_Bahasan_Gerak_Parabola

Marzano, R. J., & Kendall, J. S. (2007). *The new taxonomy of educational objectives* (2nd ed.). Corwin Press.

Mešić, V., Neumann, K., Aviani, I., Hasović, E., Boone, W. J., Erceg, N., Grubelnik, V., Sušac, A., Glamočić, D. S., Karuza, M., Vidak, A., Alihodžić, A., & Repnik, R. (2019). Measuring students' conceptual understanding of wave optics: A Rasch modeling approach. *Physical Review Physics Education Research, 15*, 1-20. https://doi.org/10.1103/PhysRevPhysEducRes.15.010115

Nabilah, M., Stepanus, S. S., Hamdani, H. (2020). Analisis kemampuan kognitif peserta didik dalam menyelesaikan soal momentum dan impuls. *Jurnal Inovasi Penelitian dan Pembelajaran Fisika, 1*(1), 1-7. https://doi.org/10.26418/jippf.v1i1.41876

Rawat, A., Kumar, S., & Singh Samant, S. (2023). A systematic review of question classification techniques based on Bloom's taxonomy. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1-7. https://doi.org/10.1109/ICCCNT56998.2023.10308403

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

Syamsyiah, Z. M., & Handayani, I. (2023). Analisis kemampuan literasi numerasi siswa SMP ditinjau dari adversity quotient dan jenis kelamin: Analysis of numerical literacy ability of junior high school students in view of adversity quotient and gender. *Edumatica: Jurnal Pendidikan Matematika, 13*(2), 136-151. https://doi.org/10.22437/edumatica.v13i02.26353

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

Titova, O., Luzan, P., Davlatzoda, Q. Q., Mosia, I., & Kabysh, M. (2023). The taxonomy approach for engineering students' outcomes assessment. *Lecture Notes in Mechanical Engineering*, 380–390. Springer. https://doi.org/10.1007/978-3-031-16651-8_36

Zoechling, S., Hopf, M., Woithe, J., & Schmeling, S. (2022). Students' interest in particle physics: Conceptualisation, instrument development, and evaluation using Rasch theory and analysis. *International Journal of Science Education, 44*(15), 2353–2380. https://doi.org/10.1080/09500693.2022.2122897