Characteristic of critical thinking test instrument and analysis students' ability using Rasch model

Siti Saffanah Apriliani, Duden Saepuzaman*, Irma Rahma Suwarma

Department of Physics Education Program, Universitas Pendidikan Indonesia, Dr. Setiabudi Street No.229, Bandung, 40154, Indonesia

*Corresponding author, email: dsaepuzaman@upi.edu

Article History
Received: 12 December 2025
Revised: 29 May 2026
Accepted: 13 June 2026

Keywords
Characteristic Instrument
Critical thinking
Rasch model
Student

Abstract
This study aims to analyze the characteristics of the instrument and students' critical thinking skills on the topic of work and energy using Rasch Model. The analysis conducted includes instrument quality, instrument characteristics as indicated by person-item fit to Rasch Model, and the distribution of students' critical thinking skills. The results of the critical thinking skills test of 82 students were analyzed using Winsteps software version 3.73. The results show that the instrument used is reliable and valid, meeting the criteria for good values for both persons and items. The instrument quality analysis shows that items and persons fit Rasch Model. The Wright Map shows the pattern of students' ability distribution and the difficulty level of items, which align with the analysis of students' abilities categorized as very high, high, low, and very low, and item difficulty levels categorized as very difficult, difficult, easy, and very easy. Item Q9 is the most difficult item, and Q6 is the easiest item. The pattern of students' responses to items is shown in the Scalogram, where there are persons who provide anomalous response patterns. Overall, the instrument used has good quality for measuring students' critical thinking skills and emphasizes the importance of having a range of question difficulties.

Apriliani, S. S., Saepuzaman, D., & Suwarma, I. R. (2026). Characteristic of critical thinking test instrument and analysis students' ability using Rasch model. *Momentum: Physics Education Journal*, *10*(1), 121-134. https://doi.org/10.21067/mpej.v10i1.13455

1. Introduction

In the era of modern development, every aspect of life is experiencing rapid change and progress, from technology and information, communication, to the increasingly complex development of science (Suyono et al., 2025). Complex global problems and such fast-paced changes require individuals to have more adaptive, analytical, and reflective thinking skills in order to face global challenges. The world of education now better understands that strengthening higher-order thinking skills, especially critical thinking, is an essential intellectual ability because this skill can support students in facing global challenges by solving problems, making decisions, and building their knowledge (Herlinawati et al., 2024; Wardani & Fiorintina, 2023). Thus, education is not just about delivering material, but about how students can build their knowledge and then apply the knowledge and experiences gained during the learning process to everyday life (Olusegun, 2015).

Critical thinking, as part of higher-order thinking skills, is a comprehensive ability that not only requires the capacity to evaluate information but also enables individuals to make logical and responsible decisions through in-depth analysis, thereby producing results (Aji et al., 2024; Facione & Gittens, 2015). This reflective and rational thinking activity is used to determine individual actions (Salah, 2030). Therefore, individuals with critical thinking skills are able to connect various ideas, find solutions to problems, provide strong arguments to support their opinions, and adapt to global developments (Yohanna et al., 2023).

In the learning process, critical thinking skills can be trained in students, and at the end of the learning process, to measure students' abilities, it is necessary to use structured assessments based on clear indicators of critical thinking skills according to existing theories. The instruments used to measure students' abilities must be of good and adequate quality to obtain data that corresponds to actual facts or conditions (Yusliani & Desnita, 2022). The quality of the instruments is very important in developing and measuring students' critical thinking skills in learning. Critical thinking skills have several aspects that need to be observed, which requires instruments capable of exploring each of

doi: 10.21067/mpej.v10i1.13455

ISSN: 2548-9135

these aspects. Instruments in the form of multiple-choice test questions can be used to measure aspects of critical thinking skills and are easy to use when processing results, but require more time in the development process (Handayani et al., 2023).

In measuring students' abilities, there are two main approaches commonly used, namely Classical Test Theory (CTT) and Item Response Theory (IRT), where the Rasch Model approach is one part of IRT (Jabrayilov et al., 2016). Many studies on educational assessment suggest that the CTT approach has several limitations, including assuming a linear relationship between latent constructs and observed scores, not allowing direct estimation of true scores without strong assumptions, and the quality parameters of items being highly dependent on the sample used (Rusch et al., 2016). Unlike the CTT approach, the Rasch Model, as part of the IRT approach, is a measurement approach developed to improve measurement reliability by converting raw data into a logit scale and allowing the analysis of student abilities and item difficulties (Boone et al., 2014; Kleppang et al., 2020). Additionally, it uses a probability-based approach, which in its analysis allows for a more accurate depiction of item quality and characteristics (Linacre, 2025). Although Rasch Model analysis is widely used in various studies, there is still limited literature that utilizes the various outputs of software used for Rasch Model analysis, so that insights related to instruments and person-item interactions can be analysed more deeply. This study seeks to bridge this gap by focusing on the quality and characteristics of instruments used to measure students' critical thinking abilities.

In physics education, critical thinking skills are particularly important in the topic of work and energy because this topic involves not only mathematical calculations but also conceptual understanding and analytical reasoning. Students often experience misconceptions related to the concepts of work, kinetic energy, potential energy, and conservation of mechanical energy. Many students tend to memorize formulas without understanding the physical meaning behind the equations and the relationships between force, displacement, and energy transformation. Solving problems in work and energy also requires students to analyze situations, interpret representations, identify relevant variables, and evaluate appropriate solution strategies. Therefore, the topic of work and energy is considered suitable for assessing students’ critical thinking skills because it demands higher-order thinking processes and conceptual reasoning abilities (Ennis, 2011; Facione & Gittens, 2015; Handayani et al., 2023).

The selection of work and energy as the context of this study was also based on preliminary observations and findings from previous studies indicating that students frequently encounter difficulties in understanding the conceptual aspects of the topic. In classroom practice, students often focus on procedural problem solving and equation substitution rather than conceptual analysis and reasoning. As a result, students may obtain correct numerical answers while still holding misconceptions about the underlying physical concepts. These conditions indicate the need for an instrument capable of measuring students’ critical thinking skills specifically in the context of work and energy learning. Furthermore, previous studies have shown that physics learning requires assessment instruments that are able to measure not only factual knowledge but also students’ reasoning and analytical abilities in solving contextual problems (Aji et al., 2024; Yohanna et al., 2023).

Therefore, this study aims to examine the quality of the instruments by testing the reliability and validity of the critical thinking skills instruments used in this research, then to observe the characteristics of the question items that make up the instruments, as well as to see the distribution of students' critical thinking abilities compared to the difficulty level of the items in the instruments used related to the subject of work and energy using Rasch Model. Therefore, the research questions are: 1) How reliable and valid are the instruments used to measure students' critical thinking skills? 2) What is the quality of the question items that make up the critical thinking skills instruments? 3) How is the distribution of students' abilities compared to the item difficulty level?

2. Method

This study employed a descriptive quantitative design using Rasch Model analysis to examine the quality and characteristics of the critical thinking skills instrument and to analyze students’ critical thinking abilities on the topic of work and energy. The research subjects were 82 students, consisting of 26 male students and 56 female students in grade XII of high school at three schools in the Lembang District, West Bandung Regency, West Java, Indonesia. The sample distribution is shown in Table 1, and the school location map is shown in Figure 1. The school distribution map was included to provide information regarding the geographical distribution of the research participants. This visualization shows that the participants were drawn from different schools within the Lembang District, thereby providing variation in student characteristics and supporting the representation of the collected data.

Table 1. Sample Distribution

School	Student	N
School A	Male	15
	Female	44
School B	Male	8
	Female	5
School C	Male	3
	Female	7
Total		82

Figure 1. School Distribution Map

The data collection instrument used was a critical thinking skills instrument adapted from the critical thinking skills instrument developed by (Yenni et al., 2022), consisting of 13 multiple-choice questions on the topic of work and energy. This instrument refers to five main aspects and indicators of Ennis' critical thinking skills presented in Table 2. Examples of the critical thinking test used in this study are presented in Figure 2, specifically item number 11.

Table 2. Aspects and Indicators of Critical Thinking Skills

Aspects of critical thinking skills	Indicators of critical thinking skills	Question indicator	Number of items	Code
Elementary clarification	Focusing the question	Identify the most appropriate questions related to physics quantities	1,2,3,4,5 6	Q1, Q2, Q3, Q4, Q5, Q6
Basic support	Assess/consider correct information	Assess the correct statement	7	Q7
Inference	Summarizing the results of induction	Summing up an event to solve a problem	8,9	Q8, Q9
Advanced clarification	Giving assumptions	Provide assumptions regarding problem solving solutions	10,11	Q10, Q11
Strategy and tactics	Determine the action/strategy	Giving action in the form of arguments related to physics problems	12,13	Q12, Q13

In weightlifting, when a person is in a full squat position and lifts a load, the load can be lifted through a distance of 0.43 meters. If the same person lifts the same load from a half-squat position, a total distance is only 0.215 meters. Assuming all other variables are the same, a person in a full squat does twice as much work compared to someone in a half squat. What must a person in a half-squat position do to perform the same amount of work as a person in a full squat ….

A. reduce the work by half
B. reduce the work by two times
C. increase the work by half
D. increase the work by two times
E. increase the work by four times

Figure 2. Item Number 11

The data collection process was carried out using Google Forms through a link shared with the participants (https://forms.gle/M1xuZZqTaCbTX38y9). Participants completed the critical thinking skills test online, making the collected data easy to analyze. The use of Google Forms in assessment has a positive and effective impact for evaluation purposes and is easier to manage (Fathonah et al., 2025; Manggaberani & Darlis, 2024). All participants consented to take part in completing the instrument, and any information related to their identities was anonymized to protect the confidentiality of the research data.

Item analysis was conducted using the Rasch Model. Rasch Model is used to evaluate test instruments and provide information about the quality of the instruments (Xiao et al., 2018). Not only does it analyze the quality of the instruments used, but this analysis also provides objective and reliable results for assessing the quality of students taking the test (Daeng Naba et al., 2024; Maulana et al., 2023). The collected data were entered into Microsoft Excel and then analyzed using the Rasch Model approach with the help of Winsteps software version 3.73. The output from Winsteps is used to analyze the quality and characteristics of the instrument as well as to examine the distribution of students' critical thinking abilities compared to the difficulty level of the items.

In this study, there are several criteria to determine the reliability or quality of the instrument through Rasch Model analysis, namely person-item reliability, Cronbach's Alpha reliability, and person-item separation. Table 3 shows the categorization of reliability, separation, and Cronbach's Alpha (Sumintono & Widhiarso, 2015). In addition, validity is observed through instrument validity, namely the average values of Infit-Outfit MNSQ and Infit-Outfit ZSTD. Construct validity or the unidimensionality of the instrument ensures whether the instrument is capable of measuring what it is supposed to measure, in this case, students' critical thinking skills (Abdaziz et al., 2014). The unidimensionality criteria of the instrument are explained by the raw variance value, with categories as follows: if the value is more than 20%, it is considered acceptable; more than 40%, good; and more than 60%, very good. Meanwhile, the ideal value for unexplained variance should not exceed 15% (Handayani et al., 2023).

Table 3. Reliability and Categorize in Rasch Model Analysis

Statistics	Interval value	Category
Cronbach's Alpha	$0.8 \leq \alpha$	Very High
	$0.7 \leq \alpha < 0.8$	High
	$0.6 \leq \alpha < 0.7$	Good
	$0.5 \leq \alpha < 0.6$	Moderate
	$\alpha < 0.5$	Low
Person-Item Reliability	$0.94 \leq r$	Excellent
	$0.91 \leq r < 0.94$	Very Good
	$0.80 \leq r < 0.91$	Good
	$0.67 \leq r < 0.80$	Sufficient
	$r < 0.67$	Low
Person-Item Separation		High separation value indicates the instrument has a good quality; it can identify the group item and person

Next, to examine the characteristics of analysis instruments, Item Fit is used to determine whether an item fits the Rasch model based on three aspects: the Mean Square outfit (MNSQ) value, Z-Standard outfit (ZSTD), and point-measure correlation (PT-MEA CORR) (Boone et al., 2014). The criteria indicating whether an item fits or is suitable for the Rasch model are when it meets the range of 0.5 < MNSQ < 1.5 for mean square outfit, -2.0 < ZSTD < +2.0 for Z-standard value, and 0.4 < Pt measure corr < 0.85. An item is considered to fit the Rasch model if it meets one, two, or all three of these criteria, and the item can be retained for data collection (Nurhayati et al., 2025). Identifying whether a question item truly measures the construct (variable) being assessed is observed based on the PT-MEA CORR value. If the PT-MEA CORR value is positive, the item functions as intended to measure the variable; whereas if the value is negative, it indicates that the item does not represent the measured variable, and such items can be revised or removed (Linacre, 2025). Furthermore, through Rasch model analysis, it can be identified whether students answer items according to their level of ability and indicate whether the responses fit the Rasch model (Handayani et al., 2023). The criteria used to assess Person Fit are the same as Item Fit, which is based on the comparison of the measurement results of the Mean Square outfit (MNSQ), Z-Standard outfit (ZSTD), and point-measure correlation (PT-MEA CORR). The threshold values used in this study for MNSQ, ZSTD, PT-MEA CORR, reliability, and separation were adopted from commonly accepted criteria in Rasch measurement analysis (Boone et al., 2014; Linacre, 2025; Sumintono & Widhiarso, 2015). These criteria are widely used in educational measurement studies to evaluate item quality, response consistency, person-item interaction, and instrument performance. Therefore, the use of these threshold values in this study provides a theoretically grounded basis for interpreting the results of Rasch Model analysis.

The characteristics of the instruments were also analyzed based on item difficulty level, students' critical thinking ability, and Differential Item Functioning (DIF), which was conducted to detect potential item bias. The item difficulty level was determined from the results of the Item Measure analysis, which was grouped based on the item logit value (LVI) and divided into several categories: very difficult (LVI ≥ M+SD), difficult (M ≤ LVI < M+SD), easy (M-SD ≤ LVI < M), and very easy (LVI < M-SD). Meanwhile, the Person Measure value was used to determine the level of students' critical thinking ability based on the person logit value (LVP) and categorized into very high (LVP ≥ M+SD), high (M ≤ LVP < M+SD), low (M-SD ≤ LVP < M), and very low (LVP < M-SD) (Adams et al., 2018). DIF analysis was conducted to detect potential item bias, whether an item provides an advantage or disadvantage to certain groups even though participants from different groups have the same ability. In this study, the groups referred to were gender, i.e., male and female students. The criterion for detecting whether an item has DIF is if the probability value or PROB is less than 0.05 (Ismail et al., 2024). Furthermore, the distribution of students' abilities and item difficulty levels were evaluated through person-item mapping using the Wright Map and response pattern examination using a scalogram. These two analyses provide a comprehensive overview of the alignment between students' abilities and the difficulty characteristics of each item.

3. Results and Discussion

3.1. Instrument Reliability and Validity

3.1.1. Reliability

Analysis using Rasch Model on the reliability of the critical thinking skills instrument was obtained from the output of the Winsteps software version 3.73. Three aspects observed regarding instrument reliability are the Cronbach's alpha coefficient (KR-20), participant reliability value, and item reliability. The results of the instrument reliability analysis and separation are shown in Table 4

Table 4. Reliability of the Instrument

	Person	Item
Mean (SD)	0.43 (1.75)	0.0 (0.76)
Separation	1.72	2.36
Reliability	0.75	0.85
Cronbach Alpha	0.84	

Table 4 shows that the reliability of persons and items are 0.75 and 0.85, respectively. The Cronbach's Alpha coefficient is 0.84. Based on Table 3, the person reliability falls into the "sufficient" category as it is in the range of 0.67 ≤ r < 0.80, while the item reliability is in the "good"category, ranging from 0.80 ≤ r < 0.91, and the Cronbach's Alpha coefficient is in the "very high" category because it has a value greater than 0.80.

The good item reliability value indicates that the instrument possesses sufficient consistency in distinguishing item difficulty levels. This finding shows that the items have relatively stable measurement characteristics and can classify students according to their ability levels. Meanwhile, the sufficient person reliability suggests that students demonstrated varying response patterns during the test. This condition may occur because critical thinking skills are influenced by differences in students' conceptual understanding, reasoning abilities, and problem-solving strategies. In addition, the high Cronbach's Alpha value indicates good interaction between persons and items, showing that the instrument consistently measures the intended construct. Similar findings have also been reported in previous Rasch analysis studies in physics education, where instruments with diverse item difficulties tend to produce more stable measurements of students' higher-order thinking skills (Daeng Naba et al., 2024; Handayani et al., 2023).

3.1.2. Validity

Figure 3 shows the results of the analysis using Rasch Model regarding the validity of the instrument (Model Fit), which indicates how well the data, namely the students' responses, fit the Rasch model.

```
     SUMMARY OF 77 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE   ERROR     MNSQ  ZSTD   MNSQ  ZSTD   |
|-----------------------------------------------------------------------------|
| MEAN       7.4       13.0        .41     .74      1.00    .1    .95    .1   |
| S.D.       3.4         .0       1.50     .16       .19    .7    .38    .8   |
| MAX.      12.0       13.0       2.70    1.06      1.43   1.9   1.97   1.9   |
| MIN.       1.0       13.0      -2.72     .59       .65  -1.5    .40  -1.5   |
|-----------------------------------------------------------------------------|
| REAL RMSE    .78 TRUE SD    1.29  SEPARATION  1.66  Person RELIABILITY  .73 |
|MODEL RMSE    .75 TRUE SD    1.30  SEPARATION  1.73  Person RELIABILITY  .75 |
| S.E. OF Person MEAN = .17                                                   |
-------------------------------------------------------------------------------
     SUMMARY OF 13 MEASURED (NON-EXTREME) Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE   ERROR     MNSQ  ZSTD   MNSQ  ZSTD   |
|-----------------------------------------------------------------------------|
| MEAN      46.9       82.0        .00     .29      1.00    .0    .95   -.1   |
| S.D.       9.4         .0        .76     .01       .15   1.1    .21    .8   |
| MAX.      62.0       82.0       1.09     .32      1.26   2.0   1.23   1.0   |
| MIN.      33.0       82.0      -1.27     .27       .70  -2.0    .50  -1.8   |
|-----------------------------------------------------------------------------|
| REAL RMSE    .30 TRUE SD     .70  SEPARATION  2.36  Item   RELIABILITY  .85 |
|MODEL RMSE    .29 TRUE SD     .70  SEPARATION  2.44  Item   RELIABILITY  .86 |
| S.E. OF Item MEAN = .22                                                     |
-------------------------------------------------------------------------------
```

Figure 3. Instrument Validity

The Measured Person section shows the analysis results for non-extreme person data (77 people). This data was chosen because for extreme people, the analysis of the average Infit-Outfit values for MNSQ and ZSTD cannot be displayed because the fit statistic cannot be calculated. Extreme values occur because students answer all items correctly or incorrectly. The average Infit-Outfit MNSQ values for person are 1.00 and 0.95. Meanwhile, the average Infit-Outfit ZSTD value is the same, namely 0.1. This shows that the student's response is in accordance with the Rasch Model. Likewise, the Measured Item shows the average Infit-Outfit MNSQ values are 1.00 and 0.95, while the average Infit-Outfit ZSTD is 0.0 and -0.1. These values are within the expected range of values so that all items are declared to be in accordance (fit) with Rasch modeling.

Further analysis was conducted to see whether the instrument is capable of measuring critical thinking skills. The results of the construct validity or unidimensionality analysis of the instrument, shown in Figure 4, indicate a raw variance value of 34.5%, and since this value is more than 20%, it falls into the acceptable category. In addition, the unexplained variance values are greater than 15%, in this case, the first and second unexplained variance values (9.4% and 8.4%). The analysis results show that the instrument is proven to meet construct validity, indicating that each item consistently works in measuring the intended construct, namely critical thinking skills.

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --			Modeled
Total raw variance in observations	=	19.9	100.0%		100.0%
Raw variance explained by measures	=	6.9	34.5%		35.0%
Raw variance explained by persons	=	4.0	20.0%		20.3%
Raw Variance explained by items	=	2.9	14.5%		14.7%
Raw unexplained variance (total)	=	13.0	65.5%	100.0%	65.0%
Unexplned variance in 1st contrast	=	1.9	9.4%	14.3%	
Unexplned variance in 2nd contrast	=	1.7	8.4%	12.8%	
Unexplned variance in 3rd contrast	=	1.4	7.3%	11.1%	
Unexplned variance in 4th contrast	=	1.3	6.6%	10.0%	
Unexplned variance in 5th contrast	=	1.2	6.1%	9.3%	

Figure 4. Construct Validity

3.2. Instrument Characteristics

3.2.1. Item Fit

The analysis of the suitability of critical thinking skill items with the model is determined by three aspects: Mean Square outfit (MNSQ), Z Standard outfit (ZSTD), and point correlation (PT-MEA CORR). The results of the statistical Item Fit analysis are shown in Figure 5.

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
9	33	82	1.09	.28	1.26	2.0	1.23	.9	A .45	.57	68.8	74.4	Q9
5	44	82	.26	.28	1.02	.2	1.21	1.0	B .56	.59	74.0	75.0	Q5
1	44	82	.26	.28	1.14	1.1	1.17	.9	C .52	.59	71.4	75.0	Q1
2	55	82	-.62	.29	1.16	1.1	.99	.1	D .53	.58	71.4	78.7	Q2
7	47	82	.03	.28	1.16	1.2	1.02	.2	E .53	.59	68.8	76.1	Q7
3	57	82	-.80	.30	.90	-.6	1.07	.3	F .60	.57	84.4	79.4	Q3
4	57	82	-.80	.30	1.06	.4	.90	-.2	G .56	.57	76.6	79.4	Q4
13	37	82	.78	.28	.95	-.3	1.02	.2	f .59	.58	74.0	73.6	Q13
8	36	82	.86	.28	.97	-.2	.82	-.8	e .60	.57	67.5	73.5	Q8
10	42	82	.41	.28	.96	-.2	.93	-.3	d .60	.58	74.0	74.3	Q10
6	62	82	-1.27	.32	.88	-.7	.65	-.9	c .61	.55	83.1	81.6	Q6
12	38	82	.71	.27	.84	-1.3	.81	-.8	b .65	.58	80.5	73.7	Q12
11	58	82	-.89	.30	.70	-2.0	.50	-1.8	a .70	.57	85.7	79.7	Q11
MEAN	46.9	82.0	.00	.29	1.00	.0	.95	-.1			75.4	76.5	
S.D.	9.4	.0	.76	.01	.15	1.1	.21	.8			6.0	2.7	

Figure 5. Statistic of the Critical Thinking Instrument Item Fit

Based on Figure 5, almost all items meet the MNSQ acceptance criteria (0.5 < MNSQ < 1.5), with only item Q11 having a value exactly at 0.5. However, all items meet the acceptance criteria for ZSTD (-2.0 < ZSTD < +2.0) and PT-MEA CORR (0.4 < Pt measure corr < 0.85). Therefore, almost all items meet the three criteria for MNSQ, ZSTD, and PT-MEA CORR values. Only one item does not meet the MNSQ criteria; overall, the critical thinking test instrument items are considered fit with Rasch Model, and further analysis can be carried out. This finding aligns with previous research, which stated that if there are a few items that do not meet one of these three criteria, all items can still be considered fit, and those items are retained for data collection (Ridwan et al., 2023).

3.2.2. Item Difficulty Level

The difficulty level of the critical thinking items in this study was grouped based on the Item Logit Value (LVI). Based on the analysis results, the average item measure and standard deviation (SD) were obtained, as shown in Table 4, which are 0.00 and 0.76, respectively. The items were grouped into four difficulty level categories: very difficult (LVI ≥ 0.76), difficult (0.00 ≤ LVI < 0.76), easy (-0.76 ≤ LVI < 0.00), and very easy (LVI < -0.76). A summary of the item difficulty level analysis is shown in Table 5.

Table 5. Difficulty Level of Critical Thinking Items

Category	LVI	Item
Very difficult	LVI ≥ 0.76	Q9, Q8, Q13
Difficult	0.00 ≤ LVI < 0.76	Q12, Q10, Q1, Q5, Q7
Easy	-0.76 ≤ LVI < 0.00	Q2
Very easy	LVI < -0.76	Q3, Q4, Q11, Q6

As shown in Table 5, of the 13 critical thinking test questions, 3 items fall into the very difficult category, 5 items into the difficult category, 1 item into the easy category, and 4 items into the very easy category. This shows that the test instrument used to measure students' critical thinking has various levels of difficulty. Of the 13 items, some are at a very difficult level.

3.2.3. Differential Item Functioning (DIF)

DIF analysis was conducted to see if there is a potential item bias. In this study, DIF analysis refers to gender bias between males and females regarding the items. The DIF results were analyzed by looking at the probability values from the Winsteps software output, as shown in Figure 6. It can be seen that the probability for item Q5 is less than 0.05, which means that the item is indicated as DIF (Ismail et al., 2024). This means that item Q5 has gender bias that could benefit or disadvantage a particular group.

The presence of DIF in item Q5 may indicate that the item contains contextual or linguistic elements that are interpreted differently by male and female students. This finding suggests the importance of carefully evaluating item construction to ensure that each question measures critical thinking skills objectively without favoring a particular group. Differences in students' experiences, familiarity with problem contexts, and interpretation of statements may contribute to the emergence of item bias. Therefore, future revisions of the instrument should consider the contextual neutrality of item representation and language use to minimize potential bias and improve fairness in assessment (Ismail et al., 2024).

DIF class specification is: DIF=$S3W1

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	BETWEEN-CLASS MEAN-SQUARE	t=ZSTD	Item Number	Name
2	.1812	1	.6704	.0764	-.7498	1	Q1
2	.2394	1	.6246	.1122	-.6266	2	Q2
2	.0922	1	.7614	.0424	-.9100	3	Q3
2	.0991	1	.7529	.0474	-.8823	4	Q4
2	4.6133	1	.0317	2.0992	1.0663	5	Q5
2	.0582	1	.8093	.0288	-.9993	6	Q6
2	.0127	1	.9104	.0056	-1.2742	7	Q7
2	.0046	1	.9460	.0020	-1.3829	8	Q8
2	2.2880	1	.1304	.8663	.3723	9	Q9
2	.5954	1	.4404	.2481	-.3169	10	Q10
2	.2935	1	.5880	.1405	-.5472	11	Q11
2	.1542	1	.6945	.0602	-.8184	12	Q12
2	.0546	1	.8153	.0208	-1.0665	13	Q13

Figure 6. Results of DIF Probability Analysis for Each Item

Figure 7 shows a further analysis of bias direction based on gender. The graph pattern indicates that Item Q5 is biased more in favor of females than males, as seen from the pattern of points on the graph showing higher DIF values for females on item Q5.

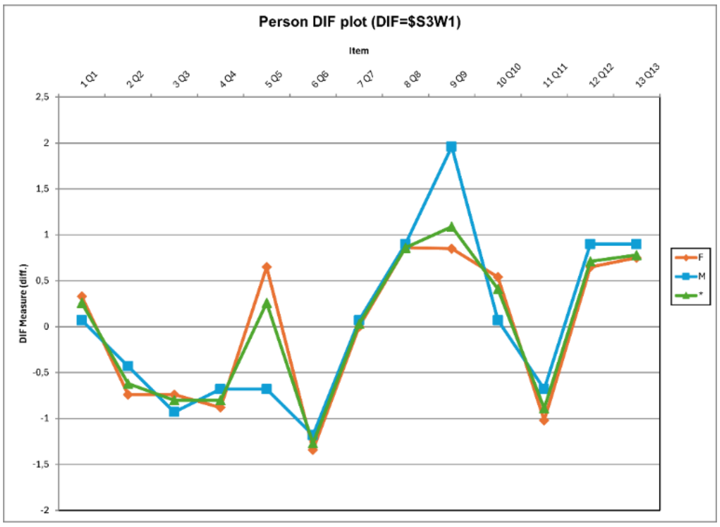

Figure 7. DIF Analysis Based on Gender

3.2.4. Person Fit

The next Rasch Model analysis is Person Fit, which is conducted to assess the consistency of students' responses to the pattern expected by the model. Similar to Item Fit, students' response patterns to Rasch Model are determined by three aspects, namely the Mean Square outfit (MNSQ) value, Z Standard outfit (ZSTD), and point-measure correlation (PT-MEA CORR). Figure 8 shows an excerpt of the Person Fit analysis results displaying the values of these three aspects.

Person STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
40	13	13	3.99	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	40F
53	13	13	3.99	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	53F
68	13	13	3.99	1.85	MAXIMUM MEASURE				.00	.00	100.0	100.0	68M
2	12	13	2.70	1.06	1.06	.4	.97	.4	.10	.17	92.3	92.3	02F
5	12	13	2.70	1.06	1.06	.4	.97	.4	.10	.17	92.3	92.3	05F
12	12	13	2.70	1.06	.95	.2	.60	.0	.30	.17	92.3	92.3	12F
18	12	13	2.70	1.06	.86	.1	.46	-.2	.42	.17	92.3	92.3	18M
20	12	13	2.70	1.06	.86	.1	.46	-.2	.42	.17	92.3	92.3	20F
22	12	13	2.70	1.06	.86	.1	.46	-.2	.42	.17	92.3	92.3	22F
23	12	13	2.70	1.06	.86	.1	.46	-.2	.42	.17	92.3	92.3	23F
27	12	13	2.70	1.06	.95	.2	.60	.0	.30	.17	92.3	92.3	27F
29	12	13	2.70	1.06	1.03	.3	.84	.3	.16	.17	92.3	92.3	29M
36	12	13	2.70	1.06	.86	.1	.46	-.2	.42	.17	92.3	92.3	36M
41	12	13	2.70	1.06	.86	.1	.46	-.2	.42	.17	92.3	92.3	41F
17	11	13	1.89	.79	.79	-.3	.52	-.6	.55	.24	84.6	84.6	17F
44	11	13	1.89	.79	1.22	.6	1.70	1.0	-.15	.24	84.6	84.6	44F
45	11	13	1.89	.79	1.22	.6	1.70	1.0	-.15	.24	84.6	84.6	45F
54	11	13	1.89	.79	1.22	.6	1.70	1.0	-.15	.24	84.6	84.6	54F
57	11	13	1.89	.79	1.13	.4	1.06	.3	.08	.24	84.6	84.6	57F

Figure 8. Excerpt of Person Fit Analysis Results

Figure 8 shows several student responses that experienced misfit with the Rasch Model. This means the person gave answers that did not align with their ability compared to the ideal model. Based on a snippet of the analysis results in Figure 8, there are students who did not meet the two criteria for Person Fit MNSQ and PT MEA CORR, namely 44F, 45F, and 54F. Meanwhile, students 18M, 20F, 22F, 23F, 36M, and 41F only met the ZSTD and PT-MEA CORR aspects, while MNSQ was outside the acceptable value range. In addition, students 02F, 05F, 12F, and 57F only met the MNSQ and ZSTD aspects. Another point shown in Figure 8 is that there are students who have maximum MNSQ and ZSTD values. These values indicate that the students answered all items correctly, which is also reflected in their total scores. Conversely, students with minimum MNSQ and ZSTD values indicate that all items were answered incorrectly. Based on the analysis of all data, it was found that 23 students (persons) had response patterns that met the value ranges for all three criteria. Meanwhile, persons who did not meet one or all three aspects of MNSQ, ZSTD, and PT-MEA CORR showed response patterns that were inconsistent with the expected answer characteristics according to Rasch Model.

3.2.5. Ability Person Measure

The analysis of Person Ability Measure based on Rasch Model was conducted to examine the distribution of students' critical thinking abilities. The pattern of students' abilities distribution is shown in Figure 9.

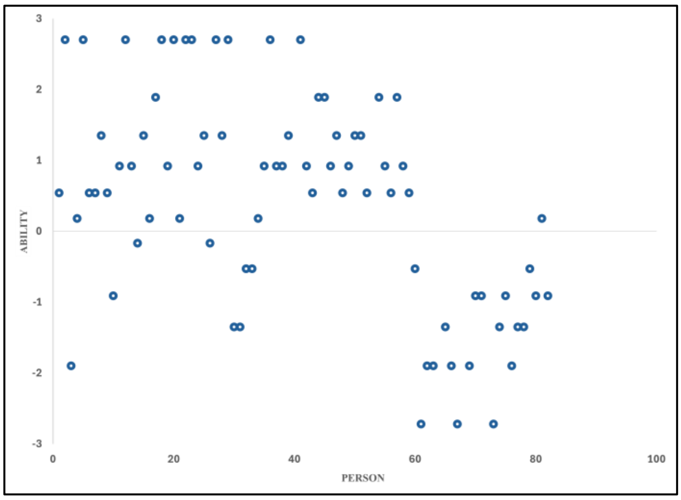

Figure 9. Distribution of Students' Critical Thinking Abilities

Figure 9 shows that the distribution of students' abilities is mostly above average. This can be seen from the distribution pattern within the zero-ability range. This distribution can be reinforced by categorizing students' ability levels based on the Logit Value Person (LVP). There are four categories of ability levels: very high (LVP ≥ 2.18), high (0.43 ≤ LVP < 2.18), low (-1.32 ≤ LVP < 0.43), and very low (LVP < -1.32). The results of students' ability levels based on this categorization are shown in Figure 10.

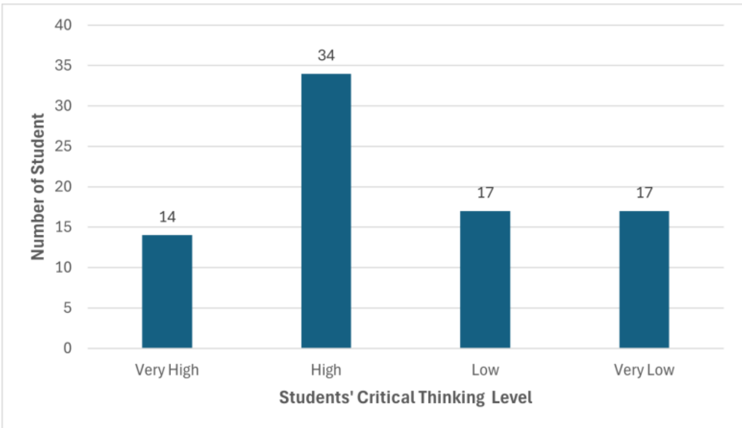

Figure 10. Categorization of Students' Critical Thinking Level

Figure 10 shows the level of critical thinking ability of 82 students based on the analysis of Person Measure with Person Logit Value (LVP). The distribution pattern of these abilities indicates a clear variation among individuals. In general, the largest proportion of students falls into the high category, with 34 students. This finding shows that most students have been able to demonstrate critical thinking skills based on the five aspects of Ennis' critical thinking. On the other hand, there are 14 students in the very high category. This number indicates that a small portion of students possesses critical thinking skills that exceed the average of their group. Meanwhile, for students in the low and very low categories, there are 17 students each.

Overall, this distribution pattern provides an overview that students' critical thinking skills are already good, but there is still a proportion of students who need to improve their abilities. Thus, this interpretation provides a basis for formulating more targeted learning recommendations to enhance students' critical thinking skills comprehensively. Physics learning should be designed to improve students' critical thinking skills through innovative learning strategies that support the development of these abilities (Suprapto et al., 2024).

3.3. Distribution of Student Ability Compared to Item Difficulty Level

3.3.1. Wright Map

Figure 11 shows a person-item comparison map produced by Winsteps software version 3.73, called the Wright Map. The right side shows the difficulty level of the items, while the left side represents the students' critical thinking abilities expressed in logit units.

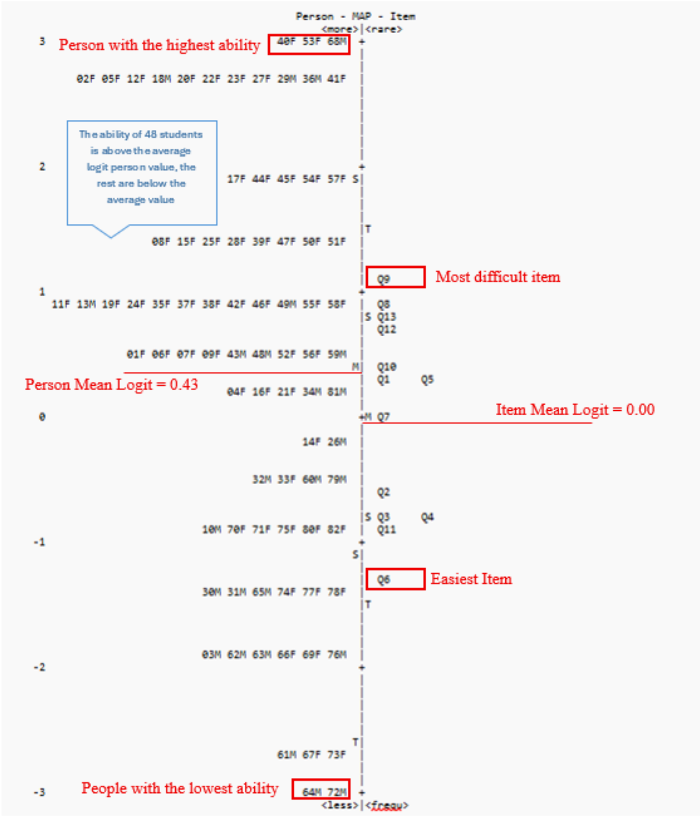

Figure 11. Wright Map of Person-Item in Critical Thinking Test

Based on the Wright Map in Figure 11, the distribution of items ranges from the easiest to the most difficult. The item difficulty levels range from -1.27 logits (item Q6, the easiest item approved by students) to 1.09 logits (item Q9, the most difficult item approved by students). Meanwhile, students with very high critical thinking abilities are positioned at the top, such as person 40F, 53F, and 68M, whereas students with very low critical thinking abilities are at the bottom, such as person 64M and 72M. Figure 11 shows that person 10M, 70F, 75F, 80F, and 82F, based on their abilities, have less than a 50% chance of correctly answering item Q9, a 50% chance of correctly answering item Q11, and more than a 50% chance of correctly answering item Q6.

In addition, through this Wright Map, it can be seen that 48 students have critical thinking abilities above the average person logit value. This aligns with previous results based on the categorization of student abilities in Figure 10, which shows that students with very high abilities (14 students) and high abilities (34 students), while the other students fall into the low and very low categories. Furthermore, the item difficulty levels shown by the Wright Map are also consistent with the findings in Table 5, where item Q5 falls into the very difficult category with an LVI ≥ 0.76, while item Q6 falls into the very easy category with an LVI < -0.76. The Wright Map provides information on the distribution of student abilities compared to item difficulty levels, which can give an overview of the learning that should be conducted based on students' ability levels, and also help evaluate the items in the instrument so that the measurement can be improved and effectively measure what it is intended to.

The Wright Map also indicates several gaps between the distribution of students' abilities and the spread of item difficulty levels. In particular, there are relatively limited items targeting students with medium-to-high ability levels. This condition suggests that the instrument may not optimally discriminate students within certain ability ranges. As a result, the measurement precision for students in these ranges may be less accurate. Therefore, additional items with moderate and higher logit difficulty levels are needed to better capture variations in students' critical thinking skills and improve measurement precision. Instruments with a balanced distribution of item difficulty levels are important in Rasch measurement because they can more effectively distinguish students across different ability categories (Boone et al., 2014; Linacre, 2025).

3.3.2. Scalogram

Further analysis is a Scalogram conducted to observe the general pattern of student responses from the easiest to the most difficult items. The pattern of student responses to the items is shown in Table 6.

Table 6. Responses of Several Students from the Scalogram

Person	Item 6134271502389	Person Code
18	+1111111111110	18M
28	+1111111110100	28F
11	+1111111110000	11F
24	+1111110110010	24F
35	+1111101110100	35F
49	+1111101110001	49M
43	+1011110110010	43M
59	+0111110001110	59M
26	+0110110110000	26M
71	+0101010100000	71F
80	+0001100001001	80F
77	+0000011000100	77F
3	+0000001000100	3M

Table 6 shows a snapshot of responses from several students from a diverse Scalogram analysis. Based on these results, it can be seen that the pattern of test participants' responses aligns with Rasch Model, showing students' consistent and ideal responses from the easiest items to the most difficult ones. This pattern is seen in persons 18M and 11F, who consistently answered the easiest question (Q6) to the most difficult question (Q9), answering correctly on the easy questions and incorrectly on the difficult ones. However, there are also respondents who gave unusual (anomalous) responses. For example, persons 80F, 77F, and 3M tend to answer difficult items correctly but answer easy questions incorrectly. Students' inconsistent responses to items indicate the possibility of guessing answers or being careless in taking the test. Thus, Scalogram data analysis is able to provide an overview of students' understanding of the learning material, allowing teachers to address existing issues. Additionally, this can serve as a recommendation for teachers to group students based on their responses to identify where the problems lie.

This study has several limitations. First, the number of participants was limited to students from three schools in one district, which may restrict the generalizability of the findings. Second, the instrument only focused on the topic of work and energy and used multiple-choice questions, so the measurement of critical thinking skills may not fully capture students' reasoning processes in depth. Third, the study relied solely on quantitative analysis using the Rasch Model without incorporating qualitative approaches such as interviews or think-aloud protocols. Therefore, future studies are recommended to involve larger and more diverse samples, develop instruments across broader physics topics, and combine quantitative and qualitative approaches to obtain a more comprehensive understanding of students' critical thinking skills.

4. Conclusion

The research findings show that the critical thinking skills instrument used meets the Rasch Model criteria for reliability, validity, and Cronbach's Alpha. The item reliability score falls into the good category (0.85) and the Cronbach's Alpha value is in the very high category (0.84). The Infit-Outfit MNSQ and ZSTD values for instrument validity fall within the expected range, indicating they conform to the model. Similarly, the construct validity shows that each item in the instrument measures the intended construct. Therefore, the instrument is considered reliable and valid and is capable of measuring students' critical thinking skills. The instrument characteristic analysis indicates that both items and persons fit the Rasch Model because one or all three criteria meet the expected ranges of outfit MNSQ, ZSTD, and PT-MEA CORR. Furthermore, it was found that the items have varying levels of difficulty, which are grouped into four categories. However, one item (Q5) was detected to have gender bias favoring females. This can serve as a recommendation for improvement so that all items are free from DIF and function fairly for all groups.

The distribution of students' critical thinking abilities compared to the item difficulty levels shown by the Wright Map indicates results consistent with the findings of item difficulty and Person Measures. Overall, it shows a pattern where most students have a high level of ability. Further analysis with the Scalogram shows students' responses to items ranging from very easy to very difficult. The results indicate that the items work well but require items with a more varied difficulty level in order to capture a broader distribution of students' critical thinking abilities.

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.

Acknowledgement

The authors would like to express their sincere gratitude to all parties who contributed to the completion of this research. Special appreciation is extended to the schools, teachers, and students who participated and collaborated during the data collection process.

References

Abdaziz, A., Jusoh, M. S., Amlus, H., & Salleh, T. S. (2014). Construct validity: A Rasch measurement model approach. *Journal of Applied Sciences Research, 9*(12), 7–12.

Adams, D., Sumintono, B., Mohamed, A., & Noor, N. S. M. (2018). E-learning readiness among students of diverse backgrounds in a leading Malaysian higher education institution. *Malaysian Journal of Learning and Instruction, 15*(2), 227–256. https://doi.org/10.32890/mjli2018.15.2.9

Aji, S. D., Ayu, H. D., Rosita, M., & Prawira, C. A. (2024). Hybrid discovery learning: Solutions for optimizing students' critical thinking and ICT skills on temperature fluctuations material. *Momentum: Physics Education Journal, 8*(1), 65–74. https://doi.org/10.21067/mpej.v8i1.8581

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

Daeng Naba, S., Istiyono, E., Kurniawan, A., & Adrianto, N. (2024). Development and validation of physics multiple-choice tests on the nature of physics using Rasch modelling analysis. *Jurnal Pendidikan MIPA, 25*(2), 671–685. https://doi.org/10.23960/jpmipa.v25i2.pp671-685

Ennis, R. H. (2011). *The nature of critical thinking: An outline of critical thinking dispositions and abilities*. University of Illinois.

Facione, P. A., & Gittens, C. A. (2015). *Think critically* (3rd ed.). Pearson.

Fathonah, S., Ahmadi, F., Lestari, W., Setiawan, D., & Yulianto, A. (2025). Pengembangan pembelajaran IPA terpadu berbasis keterampilan abad ke-21. *Jurnal IPA Terpadu, 9*(1), 37–46.

Handayani, Y., Rahmawati, R., & Widiasih, W. (2023). Using Rasch model to analyze reliability and validity of concept mastery test on electricity and magnetism topic. *JIPF (Jurnal Ilmu Pendidikan Fisika), 8*(2), 226–239. https://doi.org/10.26737/jipf.v8i2.3877

Herlinawati, H., Marwa, M., Ismail, N., Junaidi, J., Liza, L. O., Situmorang, D. D. B., & Situmorang, B. (2024). The integration of 21st century skills in the curriculum of education. *Heliyon, 10*(15), Article e35148. https://doi.org/10.1016/j.heliyon.2024.e35148

Ismail, I., Riandi, R., Kaniawati, I., Sopandi, W., & Supriyadi, S. (2024). Gender roles in understanding and implementing green energy technology in Indonesian schools: Rasch analysis. *Journal of Innovation and Educational Research, 4*(3), 298–314.

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Kleppang, A. L., Steigen, A. M., & Finbråten, H. S. (2020). Using Rasch measurement theory to assess the psychometric properties of a depressive symptoms scale in Norwegian adolescents. *Health and Quality of Life Outcomes, 18*(1), 1–8.

Linacre, J. M. (2025). *Winsteps® Rasch measurement computer program user's guide* (Version 5.10.4). Winsteps.com.

Manggaberani, A. A., & Darlis, A. M. (2024). The effectiveness of Google Forms in assessing and evaluating online learning outcomes: Meta-analysis study. *International Journal of Evaluation and Research in Education, 13*(10), 4561–4570.

Maulana, S., Rusilowati, A., Nugroho, S. E., & Susilaningsih, E. (2023). Implementasi Rasch model dalam pengembangan instrumen tes diagnostik. *Prosiding Seminar Nasional Pascasarjana (PROSNAMPAS), 6*(1), 748–757.

Nurhayati, N., Suhandi, A., Muslim, M., Kaniawati, I., Wahyudi, W., Misbah, M., & Darman, D. R. (2025). Assessing the knowledge and skills of prospective physics teachers in designing 4C skills-oriented learning: Rasch analysis. *Journal of Physics Education Research, 5*(1), 718–741.

Olusegun, S. (2015). Constructivism learning theory: A paradigm for teaching and learning. *Journal of Research & Method in Education, 5*(6), 66–70. https://doi.org/10.9790/7388-05616670

Ridwan, I. M., Kaniawati, I., Suhandi, A., Ramalis, T. R., Mahmudah, M., & Vilmala, B. K. (2023). Rasch modeling: Developing a critical thinking ability test for students on climate change material. *Momentum: Physics Education Journal, 7*(2), 247–258.

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2016). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Communications of the Association for Information Systems, 38*(1), 74–130.

Salah, S. E. (2010). Critical thinking: Reflection and perspective (Part II). *Inquiry: Critical Thinking Across the Disciplines, 26*(2), 5–17.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Trim Komunikata Publishing House.

Suprapto, N., Rizki, I. A., & Cheng, T.-H. (2024). Profile of students' physics critical thinking skills and prospect analysis of project-oriented problem-based learning model. *Journal of Turkish Science Education, 21*(1), 134–149.

Suyono, S., Harsono, P., Sutantyo, E., & Putranto, W. A. (2025). Critical thinking and 21st-century skills: Evaluating the effectiveness of problem-based learning models in multicultural classrooms. *International Journal of Educational Research Excellence, 4*(2), 511–521. https://doi.org/10.55299/ijere.v4i2.1478

Wardani, I. S., & Fiorintina, E. (2023). Building critical thinking skills of 21st century students through problem based learning model. *Jurnal Pendidikan Indonesia, 12*(3), 461–470. https://doi.org/10.23887/jpiundiksha.v12i3.58789

Xiao, Y., Han, J., Koenig, K., Xiong, J., & Bao, L. (2018). Multilevel Rasch modeling of two-tier multiple-choice test: A case study using Lawson's classroom test of scientific reasoning. *Physical Review Physics Education Research, 14*(2), Article 020104. https://doi.org/10.1103/PhysRevPhysEducRes.14.020104

Yenni, R. P., Siahaan, P., & Suwarma, I. R. (2022). ICARE-U learning model to improve critical thinking skills of high school students in Indonesia. *Jurnal Pendidikan MIPA, 12*(1), 1–7.

Yohanna, E., Siahaan, S., Muhammad, I., & Dasari, D. (2023). Trend of critical thinking skill researches in mathematics education in Scopus database across Indonesia: From research design to data analysis. *Journal of Mathematics Education and Learning, 6*(2), 151–161.

Yusliani, E., & Desnita, D. (2022). Validity and reliability of critical thinking instruments to measure the effectiveness of context-based physics e-module on wave materials. *Jurnal Penelitian Pendidikan IPA, 8*(1), 57–64.