

Evaluating Complex Problem-Solving Test Instrument Quality in Measuring Computational Thinking Skills in Global Warming: Rasch Model Analysis

Farika Nurfalalah¹, Judhistira Aria Utama^{2*}, Muslim³

¹⁻³Department of Master of Physics Education, Faculty of Mathematics and Natural Sciences,
Universitas Pendidikan Indonesia, Indonesia

ABSTRACT

This research aims to develop and evaluate a high-quality Complex Problem Solving (CPS) test instrument based on global warming issues to assess students' Computational Thinking (CT) skills. Global warming was chosen because of its contextual relevance in 21st-century education and its potential to promote higher-order thinking. The instrument consists of six essay questions that require students to demonstrate complex problem-solving skills, including decomposition, pattern recognition, abstraction, and algorithmic thinking. The method used in this study is Research and Development (R&D), which contains five stages: 1) analysis of needs; 2) development of the instrument; 3) validation by an expert; 4) trial instruments; and 5) data analysis using the Rasch Model. This instrument was evaluated using the Rasch model to examine the validity, reliability, and difficulty level of its items, which were analyzed using the Minister 3.80.1 software. The results showed that the instrument achieved the "very valid" categories for content and construct validity. The raw variance explained by the measure was 76.20%, classified as "specific," indicating strong empirical validity. The Cronbach's alpha (KR-20) value was 0.91, indicating the instrument's high reliability. The item reliability value was 0.94, and the person reliability was 0.89. This means that students' consistency in answering questions was good, and the instrument had high reliability. The distribution of item difficulty levels was balanced, allowing students to be distinguished by their abilities at various levels. It can be concluded that this instrument is reliable and valid. Thus, this instrument can measure students' computational thinking skills on global warming.

Recommended citation: Nurfalalah, F., Utama, J. A. U., & Muslim. (2025). Evaluating Complex Problem-Solving Test Instrument Quality in Measuring Computational Thinking Skills in Global Warming: Rasch Model Analysis. *Jurnal Pemberdayaan Masyarakat*, 4(2), 326-337.

ARTICLE INFORMATION

Keywords:

*complex problem solving (cps),
computational thinking (ct),
global warming, rasch model,
instrument test*

*Correspondent Author:

Judhistira Aria Utama

Email:

j.aria.utama@upi.edu

INTRODUCTION

Climate change, particularly global warming, has become a complex, multidisciplinary, and pressing global issue for addressing (Blue & Relations, 2023). The average temperature of the Earth has increased by 1.1°C since the pre-industrial era, triggering the melting of ice at the poles, rising sea levels, and an increasing frequency of hydrometeorological disasters such as floods and waves (IPCC, 2007). In Indonesia, data from the Meteorology, Climatology, and Geophysical Agency indicate that the maximum daily temperature in some areas has exceeded 35°C for more than 30 consecutive days (Supari et al., 2017). This condition demonstrates that global warming is not only a scientific phenomenon but also a challenge that demands the involvement of active participants in understanding, analyzing, and seeking solutions (Arifa, 2023; Nasution, 2019). Therefore, system education in the modern era is not only focused on aspects of cognitive theory but also emphasizes the importance of mastering skills at a high level, such as breaking down complex problems and thinking critically about real and global computational life issues (Nouri et al., 2020).

In curriculum education nationally and internationally, Computational Thinking (CT) is one of the essential competencies of the 21st century, as it supports the breakdown of problems across disciplines (Wing, 2006; Yadav et al., 2017). CT includes ability decomposition, introduction patterns, abstraction, and formulation algorithms that enable students to think systematically and solve problems (Sengupta et al., 2013). CT can be implemented through learning science-based and contextual approaches, such as addressing issues like global warming, which drives students to explore real solutions (Christensen, 2023). Besides, the Complex Problem Solving (CPS) approach is also necessary. However, based on searches conducted on the OECD webpage regarding PISA 2022 results, Indonesian students' computational thinking skills are still classified as low (OECD, 2023). In evaluation, Indonesian students' average mathematics score was 366 points, far below the OECD average of 472 points. According to Fadilah (2024), students with a certain level of mathematical disposition can only fulfill all indicators, such as computational, in solving problem type C4. Based on the Merdeka Curriculum and Directions Profile for Pancasila students, learning physics is expected to emphasize cognitive aspects and the ability to solve complex problems based on real-world contexts (Wahyudin et al., 2024). This signifies that the ability of students to model mathematics and use problem-solving strategies for more complex problems, which is the core of computational thinking, still needs significant improvement.

Preliminary observations at one of the state high schools in Karawang indicate that students' computational thinking abilities remain low. Based on the test, the diagnostics given to 35 students in class 11 MIPA show significant weakness. Only 14.29% of students could answer correctly on the "Abstraction" and "Algorithm" indicators, while 85.71% of students experienced errors. The "Pattern Recognition" indicator was answered correctly by only 34.29% of students, and "Decomposition" by 48.57%. The high error in indicators shows that students are not yet used to the process of thinking systematically, which becomes a foundational breakdown problem in the digital and technological context. Several studies have tried to develop instrument measurement skill levels based on global issues, but many are partial and not yet standardized. Wijnia et al. (2024) concluded that a problem-based assessment approach is more effective in increasing students' relevance and motivation to study global issues. Assessment based on context is also assessed more authentically because it can facilitate integrative skills between science and real life (Greiff et al., 2015). Therefore, development instrument evaluation is a specific measure of CT in the context of the global warming issue. To ensure the quality of instruments, the use of the Rasch approach becomes particularly relevant because it can objectively and empirically evaluate validity, reliability, and the level of difficulty of questions (Bond & Fox, 2015; Sumintono et al., 2015). The Rasch model provides superior statistics, such as fit analysis, item-person map, and identification of misfitting items, which allow for the development of instruments that are not only theoretical but also practical in the field of education.

This research aims to evaluate the quality of the CPS-based test instrument in the context of global warming, assessing the computational thinking skills of high school students. The evaluation was conducted using the Rasch approach to assess the validity, construct, reliability, and effectiveness of item questions in students with varying abilities (Hamzah et al., 2022). Rahmawati (2020) and Planinic et al. (2019) demonstrate that the Rasch Model effectively addresses the question breakdown problem in physics. Besides that, Yin et al. (2020) also found that CT can serve as a framework for comprehensively evaluating the physics of ability breakdown problems. This strengthens the urgency of the development instrument and the validity and contextual assessment based on global issues. The study's results are expected to contribute to strengthening the literacy environment, designing learning for the 21st century, and developing policy-supportive education that encourages active student participation in addressing real-world challenges.

METHOD

This research applies a Research and Development (R&D) approach (Sugiyono, 2019) to develop a computational thinking (CT) skills test instrument based on complex problem solving (CPS) for the global warming material. Design activity in the study refers to the stages of instrument development, according to Mardapi (2008), which consist of five stages. The stages are: (1) analysis needs, (2) development of the instrument, (3) validation by an expert, (4) trial instruments, and (5) data analysis using the Rasch Model. These stages are shown in Figure 1.

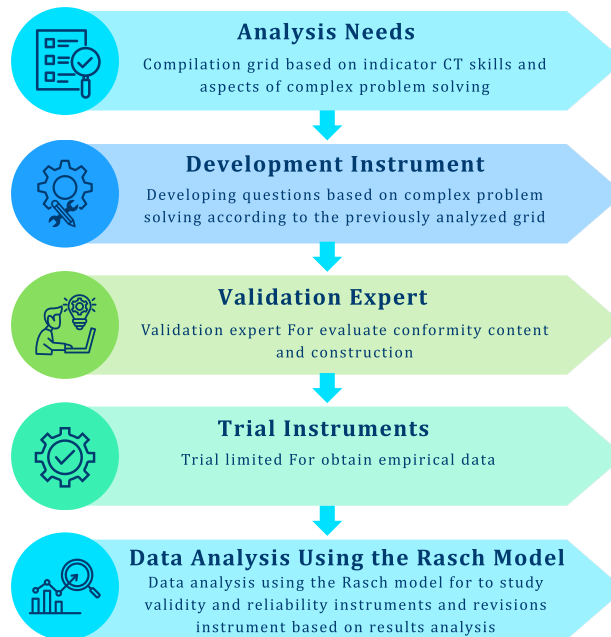


Figure 1. The stages of the research

Stages are in harmony with the development of research procedures (Boone, 2016; Mardapi, 2008). Researchers developed this instrument to measure students' computational thinking skills. This test consists of six essay questions that require complex problem-solving on the topic of global warming. Essay questions were selected to facilitate in-depth analysis of students' ability to solve problems through the stages of computational thinking. The Computational Thinking framework used to develop this test is based on Wing (2006), where the CT aspects measured consist of decomposition, pattern recognition, abstraction, and algorithms. This research focused on 30 students from Class XI of SMA Negeri 1 Jatisari Karawang, who studied material on global warming during their Physics lessons in the 2024/2025 academic year. Instruments are mainly in the form of question-shaped descriptions and sheet validation experts. Data was collected through validation content by experts and implementation tests with students to obtain responses from participants to the instrument.

Data was analyzed with two approaches. First, the validity of the content was analyzed using Aiken's V to ensure conformity of the content and construction instruments based on evaluation by experts. Second, the analysis employed the Rasch model approach using the Ministeps device software. Analysis includes item fit assessment (infit and outfit MNSQ, ZSTD), level difficulty items (logit), reliability, and index separation by item and participants, as well as person-item mapping (Wright map). The variables mainly studied are CT skills, which consist of decomposition, pattern recognition, abstraction, and algorithmic thinking (Wing, 2006).

The results of the Rasch analysis are based on revision instruments that produce valid, reliable, and appropriate measurements for assessing students' CT skills in the context of global warming. The

Rasch model was chosen in the study because of its ability to answer the five fundamental principles in measurement knowledge as put forward by Mok and Wright: (a) producing size on a linear scale; (b) accommodating missing data without bother parameters estimation; (c) provide precise parameter estimation; (d) detecting item and response participants who do not misfit; and (e) allows replication results in a way consistent (Sumintono et al., 2015).

The use of the Rasch model in this context is based on its characteristics, which allow for the measurement ability of individuals to be equal, regardless of variation from instrument measurement, and is effective in the form of tests and questionnaires (Khine, 2020). Estimating latent constructs is carried out through the analysis of the interaction between the individual's ability and the item's level of difficulty, allowing for the testing of the extent to which participants can complete every question based on the different levels of ability and complexity countered (Engelhard Jr. & Wang, 2025). Thus, the Rasch model provides a framework for powerful analytics to study validity and reliability tools measuring in the context of modern psychometrics.

RESULTS AND DISCUSSION

For instrumentation, CPS issues a global warming developed researcher used to identify students' CT ability in the school medium. In the study, the CPS tool CT diagnostics was developed by researchers to measure the ability to create and solve problems by utilizing draft-based computing, consisting of four indicators: decomposition, pattern recognition, abstraction, and algorithm. CPS issues of global warming are evaluated based on the question description. This test consists of a six-question description, divided into four customized sub-chapters, each with four CT indicators for every item question. Content validity related to level conformity between the question items and the indicator of CT skills. The study used five categories to answer determinant validity: Very Appropriate (VA), Appropriate (A), Sufficient (S), Inappropriate (I), and Very Inappropriate (VI). Analysis results of the conformity for each item in the CT test, analyzed using Aiken's V, are summarized in Table 1.

Table 1 Results of analysis of CT questions with Aiken V

No Question	Coefficient Validity of V Aiken				Category
	Material	Construction	Language	Average	
1	0.95	0.96	1.00	0.97	Very Valid
2	0.95	0.83	1.00	0.93	Very Valid
3	0.97	1.00	1.00	0.99	Very Valid
4	0.97	1.00	1.00	0.99	Very Valid
5	0.95	1.00	1.00	0.98	Very Valid
6	0.95	1.00	1.00	0.98	Very Valid

Table 1 shows the calculation of the results using Aiken V on 6 items, the question that the researcher developed. From the 6-items question, all questions are categorized as "very valid," which indicates that every item question developed by researchers has fulfilled the criteria: good materials, construction, and language. Thus, the CPS test instrument can be used to measure the CT skills of participants in education.

The items, which were the questions that fulfilled the criteria for validity of Aiken V, were further tested with 30 participants from an XI high school class that had to learn about the topic of global warming. Every participant asked for answers to all the Computational Thinking instrument test questions. The response data obtained from the trial are then analyzed through several stages of evaluation. Stages begin by testing the validity of the construct instruments, which are analyzed using the Rasch Model approach.

Validity construct refers to the extent to which a person's score on an instrument accurately represents and reflects the theoretical meaning of the construct being measured (Cronbach & Meehl, 1955). In context measurement using the Rasch Model, the validity construct can be identified through item dimensionality analysis (Sumintono & Widhiarso, 2014). The ideal instrument has several minimum dimensions, as the shown instrument can measure the construct in question in a consistent and focused manner. Item dimensionality is determined based on the raw variance value explained by the measures and the unexplained variance in the first contrast. The criteria for the feasibility of item dimensionality used in this research follow the guidelines outlined in Table 2, as proposed by Sumintono and Widhiarso (2014).

Table 2. Item Dimensionality Value Categories

Mark	Interpretation
	>20% (accepted)
Raw variance explained by measures	>40% (appropriate)
	>60% (special)
Unexplained variance 1 st contrast	Eigenvalue < 3
	Observed < 15%

This test was processed using MINISTEP software, version 5.3.3.1, and the results are presented in Figure 2.

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units				
		Eigenvalue	Observed	Expected
Total raw variance in observations	=	25.2196	100.0%	100.0%
Raw variance explained by measures	=	19.2196	76.2%	75.5%
Raw variance explained by persons	=	10.9586	43.5%	43.0%
Raw Variance explained by items	=	8.2610	32.8%	32.4%
Raw unexplained variance (total)	=	6.0000	23.8%	100.0%
Unexplned variance in 1st contrast	=	1.7532	7.0%	29.2%
Unexplned variance in 2nd contrast	=	1.2986	5.1%	21.6%
Unexplned variance in 3rd contrast	=	1.1624	4.6%	19.4%
Unexplned variance in 4th contrast	=	.9303	3.7%	15.5%
Unexplned variance in 5th contrast	=	.8160	3.2%	13.6%
Essential Unidimensionality (Rasch/Common variance) = 91.6%				

Figure 2. MINISTEP output of instrument dimensionality items, CPS test for measuring CT skills.

Figure 2 indicates that every group owns its own item statement. The raw variance is explained by a measured value of 76.20%, which falls under the "special" classification. In addition, the raw unexplained variance value of 7% indicates that the proportion of variance the instrument cannot explain is relatively small, suggesting that the items in the instrument CT test meet the assumption of unidimensionality. Based on the findings, it can be concluded that the items test has adequate validity in measuring the CT capability of educated participants. To evaluate item quality more comprehensively, an advanced analysis was carried out through empirical validity testing.

An empirical validation was conducted using the evaluation level conformity item question (fit statistic) in conjunction with the Rasch measurement model to assess the quality of the item instruments. Data for validation. This was obtained through a trial limited to the field. The validity of the analysis process is empirically determined by utilizing the Rasch Model through the MINISTEP software. One of the main menus used in the analysis is the Output Tables, specifically Table 10. Item Fit Order presents information about the compatibility of the model with the data (item fit). Assessment validity is measured using an item instrument based on three leading indicators: outfit

mean-square (MNSQ), outfit z-standard (ZSTD), and point-measure correlation (PTMEASURE-AL CORR), all presented in logit units. The criteria for eligibility of items were determined based on the limitation recommended value in the literature (Bond & Fox, 2015; Boone, 2016; Kurnianto et al., 2023; Sumintono et al., 2015), as summarized in Table 3.

Table 3. Criteria Compliance Item Question

Criteria	Value Range
MNSQ	$0.50 < \text{MNSQ} < 1.50$
ZSTD	$-2.00 < \text{ZSTD} < +2.0$
PTMEASUR-AL CORR	$0.32 < \text{PTMEASUR-AL CORR} < 0.85$

Then, each value from the third criterion is interpreted in accordance with Sumintono and Widhiarso (2014), as shown in Table 4.

Table 4. Interpretation Compliance Item Question

Condition	Interpretation
The third criterion is fulfilled.	Very Appropriate
Two of the three criteria are fulfilled	Appropriate
One of the three criteria is fulfilled	Inappropriate
No criteria fulfilled	Very Inappropriate

Table 4 presents the interpretation of determining the conformity of each item in the instrument with the Rasch measurement model. This suitability can indicate that items in the instrument's quality as a tool measure. Analysis results, in conformity with the item-by-item question test literacy science from MINISTEP, are presented in the form of an Item Statistics table. The misfit order is illustrated in Figure 2. Items that are decided to be used are presented in Table 5.

Item STATISTICS: MISFIT ORDER													
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PTMEASUR-AL CORR.	EXP.	OBS%	MATCH EXP%	Item
5	109	30	-2.29	.55	1.11	.41	2.14	1.16	A .75	.78	83.3	81.6	S5
1	105	30	-1.22	.49	1.25	.91	1.06	.30	B .77	.81	70.8	75.9	S1
3	103	30	-.75	.48	1.22	.82	.95	.08	C .83	.82	70.8	75.4	S3
6	90	30	2.03	.47	.79	-.69	.67	-.65	c .90	.87	79.2	74.5	S6
2	107	30	-1.72	.51	.78	-.68	.66	-.20	b .82	.79	79.2	77.8	S2
4	82	30	3.94	.52	.62	-1.14	.47	-.73	a .89	.86	91.7	80.2	S4
MEAN	99.3	30.0	.00	.50	.96	-.06	.99	-.01			79.2	77.6	
P.SD	9.9	.0	2.23	.03	.24	.80	.55	.64			7.2	2.6	

Figure 3 MINISTEP output item fit order instrument CT test

Table 5. Validity Results of Empirical Items in the Question CT Test

Item Question	Outfit		PT Measure Corr	Condition	Interpretation	Decision
	MNSQ	ZSTD				
1	1.06	0.30	0.77	Third criteria fulfilled	Very Appropriate	Used
2	0.66	-0.20	0.82	Third criteria fulfilled	Very Appropriate	Used
3	0.95	0.08	0.83	Third criteria, met	Very Appropriate	Used
4	0.47	-0.73	0.89	One of three criteria fulfilled	Appropriate	Nope Used
5	2.14	1.16	0.75	Two from three criteria, met	Appropriate	Used
6	0.67	-0.65	0.90	Two from three criteria, met	Appropriate	Used

Table 5 presents the results of the analysis of the conformity item question based on the Rasch model, which shows the existence of several items that do not fulfill the set criteria. Non-conformity is marked with visual indicators in the form of colored logit values (red) in both outfit parameters and point-measure correlation. Based on the analysis of the results, the Outfit mean-square (MNSQ) value is outside the recommended tolerance range for two-item questions, specifically for items numbers 4 and 5. Meanwhile, based on the z-standard outfit value (ZSTD), no item question is found that does not meet the criteria. Furthermore, in the point-measure correlation indicator (PTMEASURE-AL CORR), there are two-item questions that do not meet the criteria for conformity, namely items 4 and 6.

Based on the interpretation of the results analysis conformity item, item questions are classified into three categories, namely "Very Appropriate," "Appropriate," and "Inappropriate." Items classified in the categories "Very Appropriate" and "Appropriate" are considered worthy of use as tool measures. However, there are item questions included in the category "Very Inappropriate", namely item number 4. One item shows low performance as a tool measure, so it is stated as invalid and necessary due to instrument measurement issues. Thus, from the 6-item questions developed in the CT test, as many as five items were declared valid and can be used for the needs study.

Before being used as a measurement tool, instruments were formerly tested to identify the level of stability of the resulting scores, or what is known as reliability. The instrument can be considered reliable if it can produce accurate and consistent measurements of ability from respondents. Evaluation reliability in the study is achieved through Rasch analysis using the MINISTEP software.

In Rasch's analysis, the reliability of the instrument was reviewed based on three main parameters: person reliability, item reliability, and Cronbach's Alpha. Person reliability represents the level of consistency in answers between respondents; item reliability indicates the quality of item questions. In contrast, Cronbach's Alpha value reflects the instrument's internal consistency, which measures the consistency of responses across items (Bond & Fox, 2015). Interpretation of the person reliability and item reliability values refers to the guidelines put forward by Sumintono and Widhiarso (2015), which are presented in detail in Table 6.

Table 6. Interpretation of Person Reliability and Item Reliability Values.

Value Requirements	Interpretation
$0.94 \leq \text{Value}$	Very Good
$0.90 \leq \text{Value} < 0.94$	Good
$0.80 \leq \text{Value} < 0.90$	Enough
$0.67 \leq \text{Value} < 0.80$	Poor
$\text{Value} < 0.67$	Very Poor

Meanwhile, the Cronbach alpha values (Sumintono & Widhiarso, 2015) are provided in Table 7 for interpretation.

Table 7. Interpretation of Reliability Test Based on Cronbach Alpha Value (α)

Value Requirements	Interpretation
$0.80 \leq \alpha$	Very Good
$0.70 \leq \alpha < 0.80$	Good
$0.60 \leq \alpha < 0.70$	Enough
$0.50 \leq \alpha < 0.60$	Poor
$\alpha < 0.50$	Very Poor

Analysis results from the reliability question test literacy science, as presented by Ministep, are presented in the form of output tables. Summary Statistics are presented in Figure 4.

SUMMARY OF 30 MEASURED (EXTREME AND NON-EXTREME) Person								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD
MEAN	19.9	6.0	5.53	1.24				
SEM	.7	.0	.77	.08				
P.SD	3.7	.0	4.14	.43				
S.SD	3.8	.0	4.21	.43				
MAX.	24.0	6.0	10.81	2.03				
MIN.	9.0	6.0	-7.70	.84				
REAL RMSE	1.38	TRUE SD	3.90	SEPARATION	2.84	Person RELIABILITY	.89	
MODEL RMSE	1.31	TRUE SD	3.93	SEPARATION	3.00	Person RELIABILITY	.90	
S.E. OF Person MEAN = .77								
Person RAW SCORE-TO-MEASURE CORRELATION = .99								
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .91 SEM = 1.09								
STANDARDIZED (50 ITEM) RELIABILITY = .99								
SUMMARY OF 6 MEASURED (NON-EXTREME) Item								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD
MEAN	99.3	30.0	.00	.50	.96	-.06	.99	-.01
SEM	4.4	.0	1.00	.01	.11	.36	.25	.29
P.SD	9.9	.0	2.23	.03	.24	.80	.55	.64
S.SD	10.8	.0	2.44	.03	.26	.88	.60	.70
MAX.	109.0	30.0	3.94	.55	1.25	.91	2.14	1.16
MIN.	82.0	30.0	-2.29	.47	.62	-1.14	.47	-.73
REAL RMSE	.53	TRUE SD	2.17	SEPARATION	4.11	Item RELIABILITY	.94	
MODEL RMSE	.50	TRUE SD	2.17	SEPARATION	4.32	Item RELIABILITY	.95	
S.E. OF Item MEAN = 1.00								

Figure 4. MINISTEP summary statistics instrument output CT test

Based on Figure 4, the results of the reliability analysis for the instrument test of literacy science show diversity in three indicators. The person reliability value of 0.89 is classified as "very good," which indicates the consistency of the participant's responses to the items' questions. This condition can be caused by various factors, such as the participant's level of understanding of the questions, their ability to work on the questions, and whether the focus is on artistry questions.

The item reliability value reached 0.94, which is in the "very good" category. This value reflects the high internal consistency of the item questions, which were developed, and demonstrates that the instrument has adequate capability to measure CT capabilities accurately. This finding also strengthens the suspicion that the process of compiling instruments has been done systematically and structured, although it involved a limited number of participants. The Cronbach's alpha value is 0.91, which falls within the "very good" category, indicating high internal consistency of the instrument. Overall, a combination of high item reliability values and a high Cronbach's alpha supports the conclusion that this CT test has a decent level of reliability for use.

Analysis level difficulties aiming: To identify variation level difficulties in every item question and describe the distribution of questions based on the level of difficulties. In a study, analysis-level difficulties were addressed using the Rasch model approach through the MINISTEP software. Every item is completed in reference to the Output Tables: Item Measure. Determination level difficulties are based on the measured value in logit units. The output of the analysis is used to classify item questions based on logit measure values and standard deviation (SD) obtained from JMLE (Joint Maximum Likelihood Estimation) estimation, as proposed by [Sumintono and Widhiarso \(2015\)](#). The criteria for classifying the difficulty level of the question items are detailed in Table 8.

Table 8. Difficulty Level Criteria Item Question

Value Requirements	Difficulty Level
$1SD < \text{Measure}$	Very Difficult
$0 < \text{Measure} \leq 1SD$	Difficult
$-1SD \leq \text{Measure} \leq 0$	Easy
$\text{Measure} < -1SD$	Very easy

Analysis results in level difficulties item question test literacy science from MINISTEP in the form of an appearance table Item Statistics: Measure Order is presented in Figure 4.

Item STATISTICS: MEASURE ORDER														
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFINIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item	
4	82	30	3.94	.52	.62	-1.14	.47	-.73	.89	.86	91.7	80.2	S4	
6	90	30	2.03	.47	.79	-.69	.67	-.65	.90	.87	79.2	74.5	S6	
3	103	30	-.75	.48	1.22	.82	.95	.08	.83	.82	70.8	75.4	S3	
1	105	30	-1.22	.49	1.25	.91	1.06	.30	.77	.81	70.8	75.9	S1	
2	107	30	-1.72	.51	.78	-.68	.66	-.20	.82	.79	79.2	77.8	S2	
5	109	30	-2.29	.55	1.11	.41	2.14	1.16	.75	.78	83.3	81.6	S5	
MEAN	99.3	30.0	.00	.50	.96	-.06	.99	-.01			79.2	77.6		
P.SD	9.9	.0	2.23	.03	.24	.80	.55	.64			7.2	2.6		

Figure 4. Output of MINISTEP item measure instrument test CT

Based on Figure 4, the item 'difficult question' is item question number 4, while the item 'very easy question' is item question number 5. For categorizing the difficulty level of each question, the classification of P.SD results obtained from the Output Fit Order is presented in Table 9.

Table 9. Distribution of Difficulty Levels Item Question Test CT.

Logit Measure Value	Difficulty Level	Item Question	Amount Item Question
$2.23 < M$	Very Difficult	S4	1
$0 < M \leq 2.23$	Difficult	S6	1
$-2.23 \leq M \leq 0$	Easy	S1, S2, S3	3
$M < -2.23$	Very easy	S5	1

Based on Table 9, the CT test instrument developed in this study demonstrates a varied level of item difficulty distribution. Specifically, 16.6% of the items were categorized as "Very Difficult," 16.6% as "Difficult," 50% as "Easy," and 16.6% as "Very Easy." This even spread of item difficulty across all categories indicates that the instrument avoids clustering within a single difficulty level, which is essential for accurately distinguishing students' abilities. From the total of six items constructed, Rasch model analysis revealed that two items did not meet the criteria for validity and item fit, leading to their elimination. Item number 4 showed low discrimination power, particularly in assessing students' ability to determine the correct problem-solving phase based on computational thinking (CT) stages.

Meanwhile, item 5 met only two criteria (validity and difficulty level) and was thus excluded. The final instrument, therefore, consists of four validated items (items 1, 2, 3, and 6) that meet all psychometric criteria and adequately reflect CT dimensions. Each validated item integrates a global warming context with CT components, including the decomposition of climate-related data, identification of emission patterns, abstraction of problem-solving frameworks, and algorithmic procedures. This integration reflects the multidimensional nature of CT and situates it within an authentic, real-world issue—global warming, which adds to the contextual richness and relevance of the instrument.

The novelty of this study lies in developing and validating a CT test instrument specifically contextualized within global warming issues, which is currently underrepresented in existing educational assessments. While previous CT assessments often focus on generic or programming-based tasks, this instrument brings a real-world, complex problem-solving dimension that is both authentic and interdisciplinary. In the learning process, students need to gain meaningful experiences to optimally develop computational thinking skills (Fajri et al., 2019). Assessment through test instruments is one approach that can be used to measure these abilities (Ling et al., 2018). Based on the results of the previous discussion, several limitations exist in developing this instrument. First, the instrument has limited flexibility, making it less suitable for use on other topics without significant modification. Second, this instrument was only developed for students at grade X level, so it has not been able to measure computational thinking skills at other grade levels. Third, the focus of instrument development, which is limited to global warming, narrows the scope of the material and does not comprehensively represent students' abilities in the context of Physics subjects in general.

However, this study also has several advantages. First, the instrument developed has the potential to encourage students to improve their computational thinking skills. Second, global warming as a topic of study is a complex issue that naturally demands critical thinking skills, complex problem-solving, and computational thinking; therefore, it is very suitable to be used as an authentic context in instrument development. Third, the relevance of global warming to the current curriculum approach, which emphasizes issue-based and contextual learning, makes this instrument an effective

tool for educators in assessing student competencies in accordance with the direction of curriculum policy. Fourth, this instrument is beneficial for students and can also be used by educators to improve their computational thinking competencies.

In line with [Sulsilah et al. \(2023\)](#), the use of the Rasch model further increases the contribution of this research by providing objective measurements, identifying misfitting items, and allowing precise mapping of student abilities. This ensures not only better assessment quality but also greater potential for adaptive learning interventions. In terms of scientific contribution, this instrument provides a formative and summative assessment tool for educators aiming to foster 21st-century skills, particularly CT, in environmental science education. This aligns with [Maharani et al. \(2019\)](#), who state that an algorithm is one of the components of computational thinking, which, if appropriately designed, can facilitate the problem-solving process.

CONCLUSION

An instrument-based test for global warming was developed in the study using CPS. This has been proven to be valid and reliable for measuring the CT skills of high school students. Rasch Model analysis provides information-rich diagnostics, allowing educators to evaluate students' quality of Items and capabilities accurately. This instrument's potential is applied in evaluation learning based on problems, as well as a tool to help increase skills in the 21st century. Research should be continued with the implementation of instrument test diagnostics on students.

REFERENCES

- Arifa, M. F. (2023). PBL-Based PhET Simulations E-Worksheet: Improvement Students' Communication Skill on Global Warming in Merdeka Curriculum. *ACM International Conference Proceeding Series*, 172-177.
- Blue, J. (2023). Addressing Climate Change Challenges Through Multidisciplinary Collaboration. *Journal of Environmental Policy*, 2, 9-12.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.
- Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE Life Sciences Education*, 15(4).
- Christensen, D. (2023). Computational Thinking to Learn Environmental Sustainability: A Learning Progression. *Journal of Science Education and Technology*, 32(1), 26-44.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, 52(4), 281-302.
- Engelhard Jr, G., & Wang, J. (2025). *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences*. Routledge.
- Fadilah, N. (2024). *Analisis Kemampuan Computational Thinking Siswa SMP dalam Menyelesaikan Soal HOTS Ditinjau dari Tingkat Disposisi Matematis* [Tesis, Universitas Pendidikan Indonesia].
- Fajri, M., Yurniawati, & Utomo, E. (2019). Computational Thinking, Mathematical Thinking Berorientasi Gaya Kognitif pada Pembelajaran Matematika di Sekolah Dasar. *Dinamika Sekolah Dasar*, 1(1), 1-18.
- Greiff, S., Fischer, A., & Stadler, M. (2015). Assessing Complex Problem-Solving Skills with Multiple Complex Systems. *Thinking & Reasoning*, 21(3), 356-382.
- Hamzah, F. M., Rashid, M. N. A., Rahman, M. N. A., & Rasul, M. S. (2022). Evaluating the Validity and Reliability of Authentic Learning Instruments Using RASCH Model. *International Journal of Global Optimization and Its Application*, 1(3), 182-189.
- IPCC. (2007). *Report from Intergovernmental Panel on Climate Change*.
- Khine, M. S. (2020). *Rasch Measurement: Applications in Quantitative Educational Research*. Springer.

- Kurnianto, A., Sunjaya, D. K., Rinawan, F. R., & Hilmanto, D. (2023). Validation and Reliability of Lifestyle Instruments for Indonesian Adolescents with Hypertension Family History: A Rasch Model. *Open Access Macedonian Journal of Medical Sciences*, 11(B), 109–114.
- Ling, L. U., Tammie, C. S., Nasrah, N., Jane, L., & Norazila, A. A. (2018). An Evaluation Tool to Measure Computational Thinking Skills: Pilot Investigation. *Herald NAMSAS*, 1, 606–614.
- Maharani, S., Kholid, M. N., Pradana, L. N., & Nusantara, T. (2019). Problem Solving in the Context of Computational Thinking. *Infinity: Journal of Mathematics Education*, 8(2), 109–116.
- Mardapi, D. (2008). *Teknik Penyusunan Instrumen Tes dan Non Tes*. Mitra Cendekia Press.
- Nasution, I. B. (2019). Effectiveness Problem-Based Learning (PBL) with Reading Infusion Strategic to Improving Scientific Literacy for High School Students on Topic Global Warming. *Journal of Physics: Conference Series*, 1280(5).
- Nouri, J., Zhang, L., Mannila, L., & Norén, E. (2020). Development of Computational Thinking, Digital Competence and 21st Century Skills When Learning Programming in K-9. *Education Inquiry*, 11(1), 1–17.
- OECD. (2023). *PISA 2022 Technical Report: 6. Sample Design*.
- Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch Analysis in Physics Education Research: Why Measurement Matters. *Physical Review Physics Education Research*, 15(2), 020111.
- Rahmawati, E. N. (2020). Development of E-Handout Assisted by PhET Simulation with Problem Based Learning (PBL) Model About Momentum Conservation Law and Collision to Train Students' Conceptual Understanding. *Journal of Physics: Conference Series*, 1440(1).
- Sengupta, P., Kinnebrew, J. S., Basu, S., Biswas, G., & Clark, D. (2013). Integrating Computational Thinking with K-12 Science Education Using Agent-Based Computation: A Theoretical Framework. *Education and Information Technologies*, 18(2), 351–380.
- Sugiyono. (2019). *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. Alfabeta.
- Sulsilah, H., Hidayat, A., & Ramalis, T. R. (2023). Analysis of Computational Thinking Instrument for High School Student Using Rasch Model. *Jurnal Penelitian Pendidikan IPA*, 9(3), 1445–1450.
- Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial*. Trim Komunikata.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*.
- Supari., Tangang, F., Juneng, L., & Aldrian, E. (2017). Observed Changes in Extreme Temperature and Precipitation Over Indonesia. *International Journal of Climatology*, 37(4), 1979–1997.
- Wahyudin, D., Subkhan, E., Malik, A., Hakim, M. A., Sudiapermana, E., Alhapip, M. L., Amalia, L. S., Ali, N. B. V., & Krisna, F. N. (2024). *Kajian Akademik Kurikulum Merdeka*. Kemendikbud.
- Wijnia, L., Noordzij, G., Arends, L. R., Rikers, R. M. J. P., & Loyens, S. M. M. (2024). The Effects of Problem-Based, Project-Based, and Case-Based Learning on Students' Motivation: A Meta-Analysis. *Educational Psychology Review*, 36(1).
- Wing, J. M. (2006). Computational Thinking. *Communications of the ACM*, 49(3), 33–35.
- Yadav, A., Good, J., Voogt, J., & Fisser, P. (2017). Computational Thinking as an Emerging Competence Domain. In *Technical and Vocational Education and Training* (pp. 1051–1067). Springer.
- Yin, Y., Hadad, R., Tang, X., & Lin, Q. (2020). Improving and Assessing Computational Thinking in Maker Activities: The Integration with Physics and Engineering Learning. *Journal of Science Education and Technology*, 29(2), 189–214.