GeoScienceEd *7(3) (2026)*

Jurnal Pendidikan, Sains, Geologi, dan Geofisika

http://jpfis.unram.ac.id/index.php/GeoScienceEdu/index

Analysis of Instrument Quality for Measuring High School Students' Conceptual Understanding of Global Warming

Mohd Khairul Nizam Syah[1*], Parlindungan Sinaga[1], Dedi Sasmita[1], Lina Aviyanti[1], Rapi Alwi Azhar[1]

[1] Department of Physics Education, Faculty of Mathematics and Natural Sciences Education, Universitas Pendidikan Indonesia, Bandung, Indonesia.

DOI: https://doi.org/10.29303/goescienceed.v7i3.2453

Article Info:
Received : May, 28 2026
Revised : June, 16 2026
Accepted : June, 27 2026
Published : July, 06 2026

Correspondence:

Mohd Khairul Nizam Syah

Phone: +6283185740582

Abstract: Conceptual understanding of global warming is essential for enabling students to explain environmental phenomena scientifically and apply physics concepts to real-world issues; therefore, a valid and reliable assessment instrument is needed. This study aimed to analyze the quality of a test instrument measuring high school students' conceptual understanding of global warming using the Rasch Model. A descriptive quantitative approach was employed involving 43 eleventh-grade students from a public high school in Bandung selected through purposive sampling. The instrument consisted of 20 multiple-choice items based on the Revised Bloom's Taxonomy at the C2 level. Data were analyzed using MINISTEP. The results showed that the instrument satisfied the unidimensionality assumption, with 50.9% raw variance explained by measures and 6.7% unexplained variance in the first contrast. Item fit analysis identified 9 highly suitable items, 7 suitable items, 1 less suitable item, and 3 unsuitable items. Reliability analysis yielded person reliability of 0.63, item reliability of 0.93, and Cronbach's alpha of 0.71. Therefore, the instrument is generally suitable for measuring students' conceptual understanding of global warming, although several items require revision

Keywords: Rasch Model; Concept Comprehension Instrument; Global Warming; Validity; Reliability

Citation: Syah, M. K. N., Sinaga, P., Sasmita, D., Aviyanti, L., & Azhar, R. A. (2026). Analysis of Instrument Quality for Measuring High School Students' Conceptual Understanding of Global Warming. *Jurnal Pendidikan, Sains, Geologi, Dan Geofisika (GeoScienceEd Journal)*, 7(3), 3190–3198. https://doi.org/10.29303/goescienceed.v7i3.2453

Introduction

Global warming is one of the environmental issues that students need to understand because it is closely linked to climate change, rising global temperatures, changing weather patterns, and various ecological impacts on daily life. In physics education, the global warming curriculum does not merely emphasize students' ability to memorize information but also requires them to understand concepts, explain cause-and-effect relationships, interpret natural phenomena, and connect scientific concepts to real-world problems. Global warming is a crucial topic in physics education because it relates to environmental phenomena that are closely tied to daily life. Therefore, instruction on this topic must be supported by appropriate strategies to help students more easily grasp the concepts being taught (Rahmawati, 2023). Consequently, conceptual understanding is a vital aspect that must be developed and accurately assessed in science education.

Conceptual understanding can be defined as students' ability to grasp the meaning of a concept, connect that concept to specific phenomena, and use it to explain events scientifically. (Wandi, Mardianti, Suwarma, & Liliawati, 2023) state that conceptual understanding is one of the main aspects in the study of physics education because it relates to students' ability to construct structured meanings of physics concepts.

In the context of global warming, conceptual understanding is essential so that students not only know the definition but are also able to understand the

Email: mohdkhairulnizam@upi.edu

causes, impacts, processes involved, and mitigation efforts.(Khatmani, Liliawati, & Imansyah, 2024).demonstrate that understanding of physics concepts can be analyzed using test instruments with the aid of the Rasch Model, thereby allowing for a more objective assessment of students' abilities.

However, the results of a preliminary study conducted at a public high school in West Java Regency indicate that students' conceptual understanding of global warming is still not optimal. This preliminary study utilized a conceptual understanding test on global warming consisting of 15 multiple-choice items and involved 31 students. The analysis results showed that 19 students (61.9%) fell into the very low category, 4 students (12.9%) into the low category, 4 students (12.9%) into the moderate category, and 4 students (12.9%) into the high category. Furthermore, no students reached the very high category. These results indicate that the majority of students still face difficulties in understanding the concept of global warming.

These findings indicate that the assessment of conceptual understanding must be supported by high-quality instruments. Validity and reliability are crucial aspects in instrument development to ensure that the data obtained are empirically valid (Aini, Evendi, Halim, Syukri, & Yusrizal, 2023). Instrument quality is determined not only by the alignment of items with content indicators but also by empirical test results demonstrating that the instrument can generate valid and reliable data regarding students' abilities ((Rasyid, Istiyono, & Gunawan, 2025) and (Amelia, Muslim, & Chandra, 2020)applied Rasch analysis to examine the characteristics of a non-test instrument on sustainability awareness regarding global warming for high school students, demonstrating the relevance of the Rasch Model for evaluating global warming-related instruments.

The Rasch Model is one approach that can be used to examine the quality of an instrument in greater depth. Through this approach, researchers can obtain information regarding the instrument's validity, reliability, item difficulty levels, and the fit of test items to the measurement model(Bond & Fox, 2015) (Amiruddin et al., 2023) used the Rasch Model in a pilot study to test the validity and reliability of an instrument on global warming material and demonstrated that Rasch analysis can map instrument quality and item difficulty levels through the Wright Map.

The application of the Rasch Model to global-warming- and climate-change-related instruments has since been extended to several adjacent constructs. Ridwan et al. (2023) used Rasch analysis to develop a critical-thinking test on climate change material for physics education students, Illahi et al. (2023) applied the Rasch Model to map senior high school students' STEM literacy on global warming issues across data, technology, and human literacy dimensions, and Susongko, Ratu, & Hayati, (2021) used the Rasch Model to measure junior high school students' environmental literacy through a higher-order-thinking-skills test. These studies confirm that the Rasch Model is well established as a diagnostic tool for assessing instrument quality in environmental and science-related contexts, yet each targets a construct distinct from conceptual understanding itself, namely critical thinking, broad STEM literacy, or general environmental literacy, rather than students' conceptual mastery of global warming content specifically.

To date, no published study has specifically examined the psychometric quality, encompassing validity, reliability, and item difficulty mapping through the Wright Map, of an instrument designed to measure senior high school students' conceptual understanding of global warming. This gap is concerning in light of national-level findings showing that nearly half of Indonesian adolescents (49.7%) possess poor climate change literacy (Martha, Besral, Zainita, Rilfi, & Aminudin, 2025) which signals an urgent need for psychometrically robust assessment tools at the senior-high level capable of both certifying instrument quality and diagnosing which specific concepts students struggle with. Therefore, this study was conducted to analyze the quality of an assessment instrument for measuring high school students' conceptual understanding of global warming using the Rasch Model, encompassing instrument validity, reliability, and item difficulty levels.

Method

This study employs a descriptive quantitative approach aimed at describing the quality of a test instrument measuring students' conceptual understanding of global warming. This approach was chosen because the study focuses on a systematic analysis of the instrument's quality based on empirical data obtained through direct testing of students.

The subjects of this study were 11th-grade students at a public high school in Bandung. The research sample was determined using purposive sampling, taking into account the relevance of global warming material to the 11th-grade level. According to Campbell et al. (2020), purposive sampling is used to select respondents deemed most capable of providing information that is appropriate and beneficial to the research. Based on this technique, a sample of 43 students was obtained.

The research instrument used was a concept comprehension test consisting of 20 items tailored to the characteristics of high school students. The instrument was designed to measure the extent to which students understand concepts related to global warming. The development of test items is based on the Revised Bloom's Taxonomy indicators, specifically the cognitive domain C2 (understanding), encompassing cognitive processes such as explaining, interpreting, classifying, giving examples, comparing, and drawing conclusions.

Data analysis was conducted across three main aspects. First, instrument validity analysis was performed through unidimensionality testing based on the raw variance explained by measures and unexplained variance in the first contrast, as well as item suitability assessment using the Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (Pt Measure Corr) parameters. Second, the reliability analysis included person reliability, item reliability, and Cronbach's Alpha. Third, the analysis of item difficulty levels was based on logit values in the Item Measure table and the Wright Map visualization. All analysis results were interpreted based on the criteria established by Sumintono and Widhiarso (2015).

Result and Discussion

Instrument Validity Analysis

Unidimensionality is a key aspect of instrument evaluation because it indicates the extent to which test items consistently measure a single main construct. In the context of the Rasch Model, Boone (2016) explains that Rasch analysis is used to help researchers develop instruments and monitor measurement quality more accurately.

In this study, the unidimensionality test was conducted using the MINISTEP Rasch software via the Item Dimensionality menu. The results were interpreted based on the values of raw variance explained by measures and unexplained variance in the first contrast. The interpretation criteria are presented in Table 1.

Table 1. Interpretation guidelines for unidimensionality scores

Raw Variance Explained By Measures (%)	Criteria
>60	Special
>40	Appropriate
>20	Fulfilled

(Sumintono & Widhiarso, 2015)

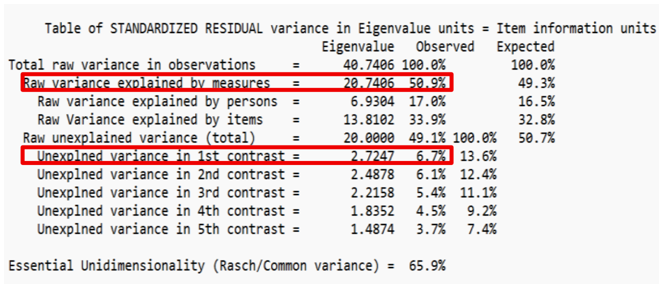
In addition to the raw variance explained by measures, the unidimensionality of the instrument can also be assessed through the unexplained variance in the first contrast. The assessment criteria are presented in Table 2. Based on the results of the unidimensionality test, the concept comprehension test instrument yielded a raw variance explained by measures of 50.9%, falling in the "Appropriate" category. The unexplained variance in the first contrast was 6.7%, which falls within the "Good" category (5–10%).

Table 2. Assessment criteria for unexplained variance in the first contrast

Unexplained Variance (%)	Criteria
> 15	Poor
10–15	Fair
5–10	Good
3–5	Very Good
< 3	Excellent

(Sumintono & Widhiarso, 2015)

These results indicate that the instrument consistently measured a single main construct, namely conceptual understanding, with no strong indication of additional dominant dimensions. Figure 1 . Dimensionality output from MINISTEP



```
     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units
                                          Eigenvalue   Observed   Expected
Total raw variance in observations     =      40.7406 100.0%         100.0%
  Raw variance explained by measures   =      20.7406  50.9%          49.3%
    Raw variance explained by persons  =       6.9304  17.0%          16.5%
    Raw Variance explained by items    =      13.8102  33.9%          32.8%
  Raw unexplained variance (total)     =      20.0000  49.1% 100.0%   50.7%
    Unexplned variance in 1st contrast =       2.7247   6.7%  13.6%
    Unexplned variance in 2nd contrast =       2.4878   6.1%  12.4%
    Unexplned variance in 3rd contrast =       2.2158   5.4%  11.1%
    Unexplned variance in 4th contrast =       1.8352   4.5%   9.2%
    Unexplned variance in 5th contrast =       1.4874   3.7%   7.4%

Essential Unidimensionality (Rasch/Common variance) =  65.9%
```

Figure 1. Raw variance explained by the measures, and the unexplained variance in the first contrast.

Furthermore, the overall validity of the instrument was reinforced through item-level analysis using three main parameters: Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (Pt Measure Corr.). The item suitability criteria are presented in Table 3.

Table 3. MNSQ, ZSTD Fit Criteria, and Pt Measure Corr

Parameter	Accepted Range
Outfit MNSQ	0.5 < MNSQ < 1.5
Outfit ZSTD	-2.0 < ZSTD < 2.0
Pt Measure Corr	0.4 < Pt Corr < 0.85

These three parameters were categorized based on item quality ratings in Table 4.

Table 4. Criteria for assessing item quality

Criteria Met	Category
All three criteria met	Very Appropriate

Two of three criteria met	Appropriate
Only one criterion met	Less Appropriate
None of the criteria met	Inappropriate

Figure 2 Item Fit Order Dimensionality output from MINISTEP

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
13	2	43	4.69	.74	1.23	.55	9.90	3.59	A-.33	.22	95.3	95.3	S13
12	3	43	4.22	.62	1.01	.18	8.12	3.40	B-.08	.26	93.0	92.9	S12
6	31	43	-.13	.39	1.00	.08	2.12	2.54	C .38	.46	79.1	77.2	S6
20	4	43	3.88	.55	1.02	.17	1.73	.98	D .17	.29	93.0	90.6	S20
7	29	43	.15	.37	1.16	.95	1.24	.84	E .34	.47	72.1	75.2	S7
2	41	43	-2.58	.76	1.19	.50	1.18	.54	F .09	.24	95.3	95.3	S2
10	32	43	-.28	.39	1.12	.66	.94	-.03	G .39	.45	72.1	78.2	S10
15	21	43	1.16	.35	1.12	.87	1.10	.51	H .40	.48	65.1	70.8	S15
16	37	43	-1.21	.48	.82	-.54	1.05	.30	I .46	.37	88.4	86.8	S16
3	35	43	-.79	.44	1.02	.18	.74	-.39	J .43	.41	79.1	83.1	S3
8	20	43	1.28	.35	.98	-.10	.96	-.11	j .49	.48	72.1	71.1	S8
4	34	43	-.61	.42	.89	-.41	.94	.03	i .48	.42	86.0	81.3	S4
5	40	43	-2.11	.63	.94	.02	.39	-.49	h .39	.29	93.0	93.0	S5
14	25	43	.67	.35	.94	-.36	.86	-.56	g .54	.48	74.4	72.2	S14
1	36	43	-.99	.46	.89	-.35	.53	-.80	f .51	.39	83.7	84.9	S1
9	29	43	.15	.37	.88	-.69	.72	-.91	e .57	.47	76.7	75.2	S9
18	39	43	-1.75	.56	.79	-.44	.36	-.72	d .50	.32	93.0	90.8	S18
17	32	43	-.28	.39	.74	-1.37	.52	-1.38	c .65	.45	81.4	78.2	S17
19	42	43	-3.35	1.03	.74	-.01	.12	-.60	b .37	.18	97.7	97.7	S19
11	40	43	-2.11	.63	.69	-.60	.25	-.79	a .51	.29	93.0	93.0	S11
MEAN	28.6	43.0	.00	.52	.96	-.03	1.69	.30			84.2	84.1	
P.SD	12.3	.0	2.14	.17	.15	.56	2.50	1.36			9.6	8.8	

Figure 2. Results of the validity analysis of the concept comprehension instrument

A summary of item validity results is presented in Table 5. The analysis results indicate that the quality of the test items varies according to their level of appropriateness in measuring students' conceptual understanding. A total of nine items (S1, S3, S8, S9, S14, S15, S16, S17, and S19) were classified as Very Appropriate. These items demonstrated excellent characteristics and were considered highly effective in assessing students' conceptual understanding. In addition, seven items (S2, S4, S7, S10, S11, S18, and S20) were categorized as Appropriate. These items met the required quality criteria and were considered suitable for use in the assessment process. Meanwhile, one item (S5) was classified as Less Appropriate, indicating that it requires improvement in several aspects. Furthermore, three items (S6, S12, and S13) were categorized as Inappropriate because they did not meet the expected quality standards.

These findings suggest that the identified items require further review and comprehensive revision before being administered in future assessments. Overall, the analysis demonstrates that the majority of the test items are appropriate for measuring students' conceptual understanding. Therefore, revising the less appropriate and inappropriate items is necessary to improve the validity, reliability, and overall quality of the assessment instrument.

Table 5. Construct Validity Analysis of the Conceptual Understanding Test Instrument

Item	**MNSQ**	**ZSTD**	**Pt Measure Corr**	**Category**
S1	0.53	-0.80	0.51	Very Appropriate
S2	1.18	0.54	0.09	Appropriate
S3	0.74	-0.39	0.43	Very Appropriate
S4	0.94	0.03	0.38	Appropriate
S5	0.39	-0.49	0.39	Less Appropriate
S6	2.12	2.54	0.38	Inappropriate
S7	1.24	0.84	0.34	Appropriate
S8	0.96	-0.11	0.49	Very Appropriate
S9	0.72	-0.91	0.57	Very Appropriate
S10	0.94	-0.03	0.39	Appropriate
S11	0.25	-0.79	0.51	Appropriate
S12	8.12	3.40	-0.08	Inappropriate
S13	9.90	3.54	-0.33	Inappropriate
S14	0.86	-0.56	0.54	Very Appropriate
S15	1.10	0.51	0.40	Very Appropriate

S16	1.05	0.30	0.46	Very Appropriate
S17	0.52	-1.38	0.65	Very Appropriate
S18	0.36	-0.72	0.50	Appropriate
S19	0.72	-0.91	0.57	Very Appropriate
S20	1.73	0.98	0.17	Appropriate

Reliability Analysis

Instrument reliability indicates the degree of consistency in the scores produced by a test (Johnson & Christensen, 2020). In this study, reliability testing was analyzed using Rasch modeling with MINISTEP software. Reliability indicators include person reliability, item reliability, and Cronbach's Alpha. The interpretation criteria are presented in Tables 6 and 7.

Table 6. Range of Person Reliability and Item Reliability Values

Range	Interpretation
> 0.94	Excellent
0.91–0.94	Very Good
0.81–0.90	Good
0.67–0.80	Fair
< 0.67	Poor

(Sumintono & Widhiarso, 2015)

Table 7. Cronbach's Alpha Range

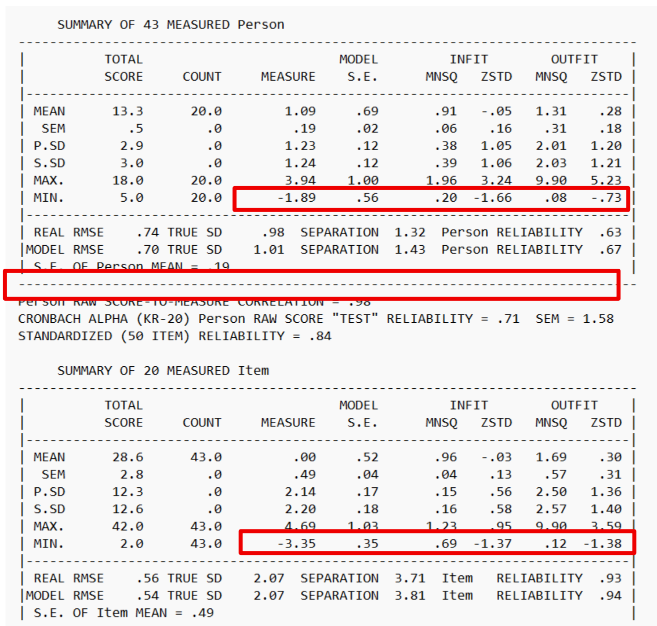
Range	Interpretation
> 0.80	Excellent
0.71–0.80	Good
0.61–0.70	Fair
0.50–0.60	Poor
< 0.50	Bad

(Sumintono & Widhiarso, 2015)

Based on the results of the Rasch model reliability analysis, the person reliability coefficient was 0.63 (Poor/adequate category), the item reliability coefficient was 0.93 (Very Good category), and Cronbach's Alpha was 0.71 (Good category). This indicates that the consistency of students' abilities falls into the adequate category, the quality of the test items falls into the very good category, and the overall reliability of the instrument falls into the good category. A person reliability value of 0.63 indicates that the instrument is sufficiently capable of distinguishing students' abilities, though the separation between high- and low-ability students is not yet very strong. (Bintang & Suprananto, 2024) explain that sample size, test length, and the alignment between respondents' ability and item difficulty levels can influence separation reliability in the Rasch Model.

Meanwhile, an item reliability coefficient of 0.93 indicates that the difficulty hierarchy of the items is consistent and reliably structured, enabling reliable measurement information.(Boone, 2016) emphasizes that Rasch analysis is beneficial in instrument development because it provides information regarding item quality, item fit, and measurement reliability. A Cronbach's Alpha of 0.71 indicates adequate internal consistency. Tavakol and Dennick (2011) explain that Cronbach's Alpha measures the extent to which the items in a test are interrelated in measuring the same construct.

Figure 3 Summary Statistics output from MINISTEP



```
     SUMMARY OF 43 MEASURED Person
 -------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT      |
|          SCORE      COUNT    MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD   |
|-------------------------------------------------------------------------------|
| MEAN      13.3       20.0       1.09     .69       .91   -.05   1.31    .28   |
|  SEM        .5         .0        .19     .02       .06    .16    .31    .18   |
| P.SD       2.9         .0       1.23     .12       .38   1.05   2.01   1.20   |
| S.SD       3.0         .0       1.24     .12       .39   1.06   2.03   1.21   |
| MAX.      18.0       20.0       3.94    1.00      1.96   3.24   9.90   5.23   |
| MIN.       5.0       20.0      -1.89     .56       .20  -1.66    .08   -.73   |
|-------------------------------------------------------------------------------|
| REAL RMSE    .74 TRUE SD     .98  SEPARATION  1.32  Person RELIABILITY  .63   |
|MODEL RMSE    .70 TRUE SD    1.01  SEPARATION  1.43  Person RELIABILITY  .67   |
| S.E. OF Person MEAN = .19                                                     |
 -------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .71  SEM = 1.58
STANDARDIZED (50 ITEM) RELIABILITY = .84

     SUMMARY OF 20 MEASURED Item
 -------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT      |
|          SCORE      COUNT    MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD   |
|-------------------------------------------------------------------------------|
| MEAN      28.6       43.0        .00     .52       .96   -.03   1.69    .30   |
|  SEM       2.8         .0        .49     .04       .04    .13    .57    .31   |
| P.SD      12.3         .0       2.14     .17       .15    .56   2.50   1.36   |
| S.SD      12.6         .0       2.20     .18       .16    .58   2.57   1.40   |
| MAX.      42.0       43.0       4.69    1.03      1.23    .95   9.90   3.59   |
| MIN.       2.0       43.0      -3.35     .35       .69  -1.37    .12  -1.38   |
|-------------------------------------------------------------------------------|
| REAL RMSE    .56 TRUE SD    2.07  SEPARATION  3.71  Item   RELIABILITY  .93   |
|MODEL RMSE    .54 TRUE SD    2.07  SEPARATION  3.81  Item   RELIABILITY  .94   |
| S.E. OF Item MEAN = .49                                                       |
 -------------------------------------------------------------------------------
```

Figure 3. Results Of The Person Reliability, Item Reliability, And Cronbach's Alpha Analyses

Difficulty Level Analysis

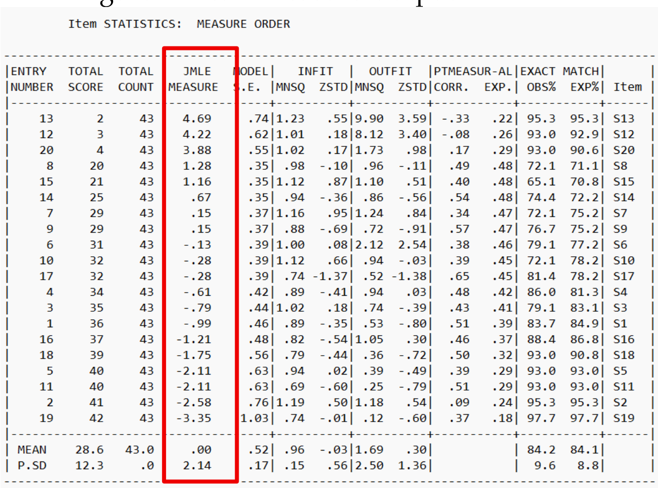
The difficulty level of each item was analyzed based on Table 13: Item Measures and Table 1: Variables (Wright Maps) outputs from MINISTEP. The criteria for interpreting item difficulty levels are presented in Table 8.

Table 8. Item Difficulty Level Criteria

Criteria	Interpretation
Measure Logit < -SD	Very Easy
-SD < Measure Logit ≤ 0.00	Easy
0.00 < Measure Logit ≤ SD	Difficult
Measure Logit > SD	Very Difficult

(Sumintono & Widhiarso, 2015)

Figure 4 – Item Measure Output


Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
13	2	43	4.69	.74	1.23	.55	9.90	3.59	-.33	.22	95.3	95.3	S13
12	3	43	4.22	.62	1.01	.18	8.12	3.40	-.08	.26	93.0	92.9	S12
20	4	43	3.88	.55	1.02	.17	1.73	.98	.17	.29	93.0	90.6	S20
8	20	43	1.28	.35	.98	-.10	.96	-.11	.49	.48	72.1	71.1	S8
15	21	43	1.16	.35	1.12	.87	1.10	.51	.40	.48	65.1	70.8	S15
14	25	43	.67	.35	.94	-.36	.86	-.56	.54	.48	74.4	72.2	S14
7	29	43	.15	.37	1.16	.95	1.24	.84	.34	.47	72.1	75.2	S7
9	29	43	.15	.37	.88	-.69	.72	-.91	.57	.47	76.7	75.2	S9
6	31	43	-.13	.39	1.00	.08	2.12	2.54	.38	.46	79.1	77.2	S6
10	32	43	-.28	.39	1.12	.66	.94	-.03	.39	.45	72.1	78.2	S10
17	32	43	-.28	.39	.74	-1.37	.52	-1.38	.65	.45	81.4	78.2	S17
4	34	43	-.61	.42	.89	-.41	.94	.03	.48	.42	86.0	81.3	S4
3	35	43	-.79	.44	1.02	.18	.74	-.39	.43	.41	79.1	83.1	S3
1	36	43	-.99	.46	.89	-.35	.53	-.80	.51	.39	83.7	84.9	S1
16	37	43	-1.21	.48	.82	-.54	1.05	.30	.46	.37	88.4	86.8	S16
18	39	43	-1.75	.56	.79	-.44	.36	-.72	.50	.32	93.0	90.8	S18
5	40	43	-2.11	.63	.94	.02	.39	-.49	.39	.29	93.0	93.0	S5
11	40	43	-2.11	.63	.69	-.60	.25	-.79	.51	.29	93.0	93.0	S11
2	41	43	-2.58	.76	1.19	.50	1.18	.54	.09	.24	95.3	95.3	S2
19	42	43	-3.35	1.03	.74	-.01	.12	-.60	.37	.18	97.7	97.7	S19
MEAN	28.6	43.0	.00	.52	.96	-.03	1.69	.30			84.2	84.1	
P.SD	12.3	.0	2.14	.17	.15	.56	2.50	1.36			9.6	8.8	

Figure 4. Difficulty Level Results Based on Item Measure Output

The classification of each test item by difficulty level is presented in Table 9. Based on the analysis, the test items were distributed across four levels of difficulty. Two items, namely S2 and S19, were classified as Very Easy. These items could be answered correctly by most students, indicating that they required only a basic level of conceptual understanding. In addition, ten items (S1, S3, S4, S5, S6, S10, S11, S16, S17, and S18) were categorized as Easy. These items were considered appropriate for assessing students' mastery of fundamental concepts while remaining accessible to the majority of learners. Furthermore, five items (S7, S8, S9, S14, and S15) were classified as Difficult, suggesting that they required higher-order thinking skills and a deeper understanding of the subject matter.

Three items, namely S12, S13, and S20, fell into the Very Difficult category. The high level of difficulty of these items may have challenged students beyond the expected level of competence. Overall, the findings indicate that the test contained a wide range of item difficulty levels, providing opportunities to differentiate students' conceptual understanding. Therefore, balancing the proportion of easy, moderate, and difficult items is recommended to improve the effectiveness and fairness of the assessment instrument.

Table 9. Classification of item difficulty levels

Item	JMLE Measure	Standard Deviation	Difficulty Level
S1	-0.99	2.14	Easy
S2	-2.58	2.14	Very Easy
S3	-0.79	2.14	Easy
S4	-0.61	2.14	Easy
S5	-2.11	2.14	Easy
S6	-0.13	2.14	Easy
S7	0.15	2.14	Difficult
S8	1.28	2.14	Difficult
S9	0.15	2.14	Difficult
S10	-0.28	2.14	Easy
S11	-2.11	2.14	Easy
S12	4.22	2.14	Very Difficult

S13	4.69	2.14	Very Difficult
S14	0.67	2.14	Difficult
S15	1.16	2.14	Difficult
S16	-1.21	2.14	Easy
S17	-0.28	2.14	Easy
S18	-1.75	2.14	Easy
S19	-3.35	2.14	Very Easy
S20	3.88	2.14	Very Difficult

Figure 5 – Wright Map

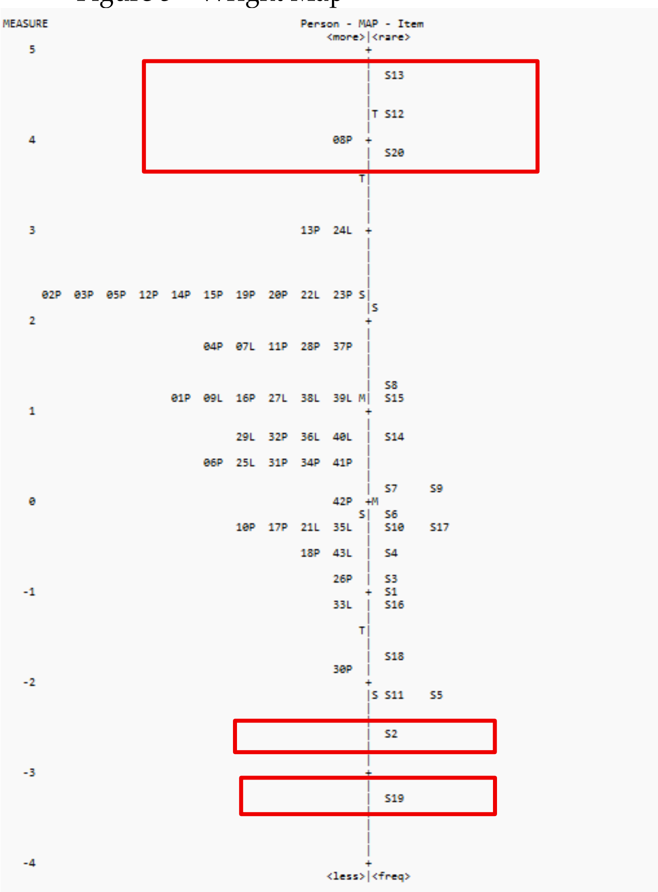


Figure 5. Wright Map

The Wright Map visually confirms these findings by showing that the instrument contains items with a broad range of difficulty levels. Items S13, S12, and S20 are located at the highest end of the logit scale, indicating that they are the most difficult items, whereas S19 and S2 are positioned at the lowest end, making them the easiest items. Most items are distributed around the middle to lower part of the scale, indicating that the instrument primarily targets students with low to moderate ability levels while still including several challenging items. (Boone, 2016) explains that Rasch analysis is useful for evaluating instrument quality by examining both item difficulty distribution and item functioning.

Discussion: Instrument Validity

The results of the unidimensionality test showed a raw variance explained by measures of 50.9%, indicating that the instrument fulfilled the assumption of unidimensionality. (Rosli et al., 2020) state that a raw variance explained by measures of at least 20% is sufficient to indicate that the instrument meets the criteria for unidimensionality, while the unexplained variance in the first contrast should not exceed 15%. The unexplained variance in the first contrast of 6.7% further confirms that there is no strong evidence of a dominant additional dimension beyond the primary construct. These results are consistent with (Loh, Jamil, & Thoe, 2022), who explain that these indicators are important in assessing the unidimensionality of an instrument using Rasch analysis.

The item fit analysis results show that the majority of test items meet the Rasch Model fit criteria. Items classified as Less Appropriate (S5) and Inappropriate (S6, S12, and S13) need to be reviewed before further use. Item inappropriateness can be caused by unclear wording, disproportionate difficulty levels, or items that do not fully reflect the construct being measured. (Rosli et al., 2020) also state that items falling outside the acceptable range of fit statistics need to be revised. This finding is in line with (Planinic, Boone, Susac, & Ivanjek, 2019) who explained that the use of Rasch analysis in physics education research plays an important role in mapping students' abilities and item difficulty levels.

Discussion: Reliability Analysis

A person reliability score of 0.63 indicates that the consistency of students' responses is at a sufficient level. (Linacre, 2012)) explains that reliability in the Rasch model relates to the repeatability of the relative positions of measures, both for respondents and test items. This can be influenced by the number of test items, the number of respondents, and the alignment of item difficulty levels with student ability. An item reliability coefficient of 0.93 indicates that the difficulty hierarchy of the test items is reliably structured. A Cronbach's Alpha of 0.71 indicates the instrument has adequate internal consistency. These results support the findings of (Erfan, Maulyda, Hidayati, Astria, & Ratu, 2020), which demonstrated that the Rasch Model can be used to examine the validity and reliability of cognitive instruments based on empirical data.

Discussion: Difficulty Level Analysis

The analysis shows that the item difficulty distribution ranged from Very Easy to Very Difficult, with a dominance of Easy items. In the Rasch Model, item difficulty is expressed in logit units, allowing item positions to be compared with students' abilities on the same scale. The fact that most items fall into the Easy category suggests that the instrument is more accessible to students at lower ability levels. (Saputri, Zulhijrah, Larasati, & Shaleh, 2023) note that item difficulty indicates the extent to which an item is considered easy or difficult for students. Items in the Very Easy and Very Difficult categories need to be reviewed to achieve a more balanced distribution of item difficulty.

Conclusion

Based on the research findings, the conceptual understanding test instrument for senior high school students on global warming material demonstrates fairly good quality based on Rasch Model analysis. The validity test results showed that the instrument fulfilled the assumption of unidimensionality, with a raw variance explained by measures of 50.9% and an unexplained variance in the first contrast of 6.7%, indicating that the instrument consistently measured students' conceptual understanding.

The item fit analysis showed that of the 20 items, 9 items were categorized as Very Appropriate, 7 items as Appropriate, 1 item (S5) as Less Appropriate, and 3 items (S6, S12, and S13) as Inappropriate. The reliability analysis yielded a person reliability value of 0.63 (adequate), item reliability of 0.93 (very good), and Cronbach's Alpha of 0.71 (good). The item difficulty analysis showed a distribution ranging from Very Easy to Very Difficult, dominated by the Easy category. Based on these findings, future researchers are advised to revise items S5, S6, S12, and S13 in terms of wording, indicator suitability, and difficulty level. Instrument trials should also be conducted with a larger and more diverse sample to strengthen the analysis results.

Acknowledgements

The authors would like to express their gratitude to the Department of Physics Education, Universitas Pendidikan Indonesia, and to all students who participated in this study.

References

Aini, E., Evendi, E., Halim, A., Syukri, M., & Yusrizal, Y. (2023). Relationship between misconceptions and students' scientific literacy abilities on global warming material. Jurnal Penelitian Pendidikan IPA, 9(10), 8051–8058. https://doi.org/10.29303/jppipa.v9i10.5156

Amelia, A., Muslim, M., & Chandra, A. F. (2020). Karakteristik instrumen non-tes sustainability awareness menggunakan analisis Rasch model materi pemanasan global untuk siswa sekolah menengah. WaPFi: Wahana Pendidikan Fisika, 5(2). https://doi.org/10.17509/wapfi.v5i2.27431

Amiruddin, M. Z. B., Samsudin, A., Suhandi, A., Kaniawati, I., Coştu, B., Aminuddin, A. H., & Kurniawan, F. (2023). Validity and Reliability of the Global Warming Instrument: A Pilot Study Using Rasch Model Analysis. Jurnal Pendidikan MIPA, 24(4), 912–922. https://doi.org/10.23960/jpmipa/v24i4.pp912-922

Bintang, R. S., & Suprananto. (2024). The Impact of Sample Size, Test Length, and Person-Item Targeting on The Separation Reliability in Rasch Model: A Simulation Study. Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia, 13(1), 1–13. https://doi.org/10.15408/jp3i.v13i1.27975

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). https://doi.org/10.1187/cbe.16-04-0148

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis Kualitas Soal Kemampuan Membedakan Rangkaian Seri Dan Paralel Melalui Teori Tes Klasik Dan Model Rasch. Indonesian Journal Of Educational Research and Review, 3(1), 11. https://doi.org/10.23887/ijerr.v3i1.24080

Johnson, & Christensen. (2020). Educational research: Quantitative, qualitative, and mixed approaches. SAGE Publications.

Khatmani, F. K., Liliawati, W., & Imansyah, H. (2024). Analisis pemahaman konsep fisika siswa smp pada materi suhu dan kalor menggunakan tes isomorfik : Rasch model. WaPFi (Wahana Pendidikan Fisika), 9(1), 11–18. https://doi.org/10.17509/wapfi.v9i1.52655

Linacre, J. M. (2012). A User's Guide to Winsteps Ministep Rasch-Model Computer Programs. Winsteps.com.

Loh, S. L., Jamil, N. A., & Thoe, N. K. (2022). Establishing the Validity and Reliability of a Program Evaluation Questionnaire using Rasch Measurement Model. Dinamika Jurnal Ilmiah Pendidikan Dasar, 14(2), 149. https://doi.org/10.30595/dinamika.v14i2.14786

Martha, E., Besral, Zainita, U. H., Rilfi, N. A., & Aminudin, S. A. (2025). Adolescents' Knowledge on Climate Change: A Nationwide Study in Indonesia. International Journal of Environmental Research and Public Health, 22(4), 571. https://doi.org/10.3390/ijerph22040571

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). https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Rahmawati, F. (2023). Meningkatkan Hasil Belajar Siswa Materi Pemanasan Global: Penggunaan Media Pembelajaran Puzzle Maker. Indonesian Journal of Action Research, 2(1), 43–48. https://doi.org/10.14421/ijar.2023.21-06

Rasyid, F., Istiyono, E., & Gunawan, C. W. (2025). Developing and analyzing items of a physics conceptual understanding test on wave topics for high school students using the Rasch Model. REID (Research and Evaluation in Education), 11(1), 1–16. https://doi.org/10.21831/reid.v11i1.75575

Ridwan, I. M., Kaniawati, I., Suhandi, A., Ramalis, T. R., Mahmudah, I. R., & 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.

Rosli, R., Abdullah, M., Siregar, N. C., Abdul Hamid, N. S., Abdullah, S., Beng, G. K., … Bais, B. (2020). Student Awareness of Space Science: Rasch Model Analysis for Validity and Reliability. World Journal of Education, 10(3), 170. https://doi.org/10.5430/wje.v10n3p170

Saputri, H. A., Zulhijrah, Larasati, N. J., & Shaleh. (2023). Analisis Instrumen Assesmen: Validitas, Reliabilitas, Tingkatkesukaran Dan Daya Pembeda. Didatik: Jurnal Ilmiah PGSD FKIP Universitas Mandiri, 09(05), 2986–2995.

Sumintono, B., & Widhiarso, W. (2015). Aplikasi Pemodelan RASCH pada. (155), 50–51.

Susongko, P., Ratu, I. M., & Hayati, M. N. (2021). Tes Kemampuan Berfikir Tingkat Tinggi untuk Mengukur Literasi Lingkungan Hidup bagi Siswa Sekolah Menengah Pertama dengan Aplikasi Model Rasch. PSEJ (Pancasakti Science Education Journal), 6(2), 93–101.

Wandi, W., Mardianti, F., Suwarma, I. R., & Liliawati, W. (2023). Theory and Practice of Conceptual Understanding in Physics Education: A Literature Review and Bibliometric Analysis of the Recent Decades. Kasuari: Physics Education Journal (KPEJ), 6(2), 107–117. https://doi.org/10.37891/kpej.v6i2.483