Integrated Science Education Journal
Vol. 7, No. 2, March 2026, pp. 255~264
ISSN: 2716-3725, DOI: 10.37251/isej.v7i2.2383

Research Article

# Development of Scientific Literacy Skills Test Instruments in Elementary School: Analysis of The Rasch Model

**Milan Nur Familia[1], Rosita Putri Rahmi Haerani[1,*], Erna Suhartini[1]**
[1] Department of Elementary Education, Mulawarman University, Kalimantan Timur, Indonesia

**Article Info**

***Article history:***

Received Jan 14, 2026
Revised Feb 21, 2026
Accepted Mar 11, 2026
OnlineFirst Mar 23, 2026

***Keywords:***

Global Warming
Instrument Assessment
Rasch Model
Scientific Literacy

## ABSTRACT

**Purpose of the study:** This study aims to develop and evaluate a context-based scientific literacy assessment instrument on the topic of global warming for sixth-grade elementary school students, employing the Rasch model to examine its validity, reliability, item difficult levels, and Differential Item Functioning (DIF).

**Methodology:** The research design followed a systematic instrument development procedure supported by Rasch model analysis. The instrument comprised 20 multiple-choice items based on the OECD scientific literacy framework and was piloted with 26 sixth-grade students. Data were analyzed using Ministep to examine validity, reliability, item difficult, and Differential Item Functioning (DIF).

**Main Findings:** The findings show that most items satisfied the validity criteria based on Outfit MNSQ, ZSTD and PTM Corr indices. Item difficulty levels were proportionally distributed from very easy to very difficult. Reliability analysis yielded a person reliability of 0.70, an item reliability of 0.72, and a Cronbach's alpha of 0.78. DIF analysis indicated that all items were free from gender bias. As a pilot study with a limited sample from one school, these results represent preliminary evidence rather than as a final validation of the instrument.

**Novelty/Originality of this study:** This study aims to develop and evaluate a scientific literacy test specifically designed for elementary school students on the topic of global warming using Rasch analysis, an area that remains limited in prior research. The instrument integrates real-life contextual situations aligned with current curriculum demands and has the potential to strengthen scientific literacy assessment practices at the elementary education level.

***Corresponding Author:***
Rosita Putri Rahmi Haerani,
Department of Elementary Education, Universitas Mulawarman,
Jl. Kuaro, Gn. Kelua, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75117, Indonesia
Email: rosita.putri.rahmi@fkip.unmul.ac.id

## 1. INTRODUCTION

Advancements in science and technology in the 21st century require individuals to master a range of essential skills, one of which is scientific literacy [1]. Scientific literacy refers to the ability to identify science-related issues, explain scientific phenomena, and use scientific evidence to propose solutions to problems [2]. Scientific literacy should be developed from an early age, particularly at the elementary school level, to enable students to progressively acquire more comprehensive competencies [3]. Other perspectives also emphasize that education at the elementary school level plays a strategic role in establishing a foundation of scientific literacy through the internalization of essential basic science concepts, which serve as the basis for the development of scientific competence at higher levels of education [4]. At the elementary school level, scientific literacy involves various science process skills, including observing, classifying objects or events, making predictions, drawing conclusions, and communicating experimental results in a systematic manner [5].

Globally, scientific literacy assessment is conducted by the Organisation for Economic Co-operation and Development (OECD) through the Programme for International Student Assessment (PISA) for 15 year old students [6]. The PISA assessment is implemented as an evaluative effort to examine the performance of education systems across countries, focusing on the measurement of students' competencies in three essential domains: reading literacy, mathematics, and science [7]. Based on the 2022 PISA results, Indonesian students obtained a scientific literacy score of 383, ranking 69th out of 81 participating countries. These results indicate that Indonesia remains below the international average and has experienced a decline compared to its performance in the 2018 PISA assessment [8]. Such findings reflect the relatively low mastery of fundamental competencies among Indonesian students in addressing real-world problems in a scientific and logical manner, suggesting that science instruction at the elementary school level has not yet fully optimized the development of students' scientific literacy skills.

Given the low scientific literacy achievement of Indonesian students, assessment instrument capable of measuring scientific literacy skills accurately and comprehensively are urgently needed. Furthermore, the Merdeka Belajar policy, which emphasizes literacy based assessment including scientific literacy requires teachers to possess professional competence in designing assessment instruments that are constructively aligned and meet established criteria of validity and reliability [9]. This view is supported Magdalena et al [10], who argue that the instruments used in assessment constitute a crucial factor in determining the quality of both learning process evaluation and learning outcomes.

Several previous studies have developed scientific literacy instruments across diverse learning contexts and educational levels. One such study was conducted Zhang et al [11], who designed a scientific literacy assessment instrument for lower and upper secondary school students, covering interdisciplinary science content, including physics, chemistry, biology, and geography. The instrument was developed based on the PISA 2015 scientific literacy framework. Another relevant study Atta et al [12] focused on the development and evaluation of a scientific literacy instrument for junior secondary school students. That study aimed to identify the scientific literacy constructs, item characteristics, and overall instrument quality through Rasch analysis using Winsteps and SPSS software.

Previous studies have predominantly focused on the development and evaluation of scientific literacy instruments for adolescent learners at the secondary education level (junior and senior high schools), whereas research addressing the development and evaluation of scientific literacy instruments at the elementary school level remains relatively limited. This condition is of particular concern, as scientific literacy should be fostered from primary education as a foundational basis for the development of scientific competencies at subsequent educational stages. Furthermore, the consistently low scientific literacy performance of Indonesian students underscores the urgent need for assessment instruments that can measure scientific literacy accurately, objectively, and in alignment with the cognitive and developmental characteristics of elementary school learners. Accordingly, this study aims to develop and evaluate a scientific literacy instrument specifically designed for Grade VI elementary school students on the topic of global warming. The global warming topic was selected because it represents a contextual environmental issue closely related to students' everyday experiences and holds strategic importance in sustainability education [13]. This context enables students to meaningfully connect scientific concepts with real-world environmental problems, thereby making it an appropriate basis for assessing scientific literacy. Although global warming has been widely employed as a context in scientific literacy instrument development, instruments explicitly designed for elementary school students on this topic remain scarce, despite its strong relevance to daily life and its significance in fostering early scientific literacy and sustainability awareness.

As an effort to address the identified research problem, this study developed a context-based scientific literacy instrument on the topic of global warming and evaluated it using Rasch model analysis. The Rasch approach was selected because it enables the separate estimation of students' abilities and item difficulty levels, while providing objective information regarding item fit and overall instrument quality. This approach is particularly relevant to the heterogeneous ability profiles commonly found among elementary school students. The scientific literacy framework adopted in this study is grounded in the PISA 2025 framework and focuses on three core competencies: (1) explaining scientific phenomena; (2) designing and evaluating scientific investigations and critically interpreting scientific data and evidence; and (3) investigating, evaluating, and using scientific information as a basis for decision-making and action.

## 2. RESEARCH METHOD

The research design followed a systematic instrument development procedure supported by Rasch model analysis, enabling the measurement results to be expressed on an interval scale and to remain consistent when applied across different groups of respondents [14]. The instrument development procedure was conducted through several stages, beginning with the preparation of test items that were developed entirely from scratch based on the learning outcomes of the Merdeka Curriculum and the components of scientific literacy. The initial set of items was then reviewed and validated by three experts to ensure content validity, construct validity, linguistic clarity, and alignment with scientific literacy principles. Following expert revision, the refined items were administered in a limited trial. The instrument consisted of 20 multiple-choice items developed in accordance with the scientific literacy competencies outlined in the OECD 2025 framework, as presented in Table 1.

Table 1. Specification Matrix of the Scientific Literacy Instrument

| Scientific Literacy Competencies | Scientific Literacy Indicators | Question Item |
|---|---|---|
| Explaining phenomena scientifically | Elaborate on how scientific knowledge may influence societal conditions and development. | 3 |
| | Making and proving accurate scientific predictions and solutions. | 15, 14, 10, 5, 7 |
| | Identify and formulate well defined hypotheses regarding observable phenomena in the natural world. | 12, 4 |
| | Remembering and applying appropriate scientific knowledge. | 2, 11 |
| Construct and evaluate designs for scientific enquiry and interpret scientific data and evidence critically | Propose an appropriate experimental design. | 9 |
| | Identifying questions in a given scientific study. | 18, 19 |
| | Interpret information displayed through various forms of representation, formulate sound conclusions grounded in the data, and critically appraise the respective merits of each format. | 6 |
| | Evaluating an experimental design that is appropriate to answer research questions. | 20 |
| Research, evaluate and use scientific information for decision making and action | Providing justification for decisions through scientifically grounded reasoning, whether individually or collaboratively, to address contemporary challenges or support sustainable development. | 17, 8, 16, |
| | Differentiate assertions supported by robust scientific evidence from those based on expertise, non-expert views, or mere opinion, and articulate the rationale for these distinctions. | 13 |
| | Building arguments to support proper scientific conclusions from a set of data. | 1 |

The sample in this study consisted of 26 sixth grade elementary school students. This sample size indicates that the study was conducted as a pilot study or limited trial, aimed at examining the feasibility, clarity, and preliminary quality of the instrument, rather than drawing definitive or generalizable conclusions.

The feasibility of the test items was evaluated using Rasch model analysis with the assistance of Ministep software. This analysis was conducted to assess the quality of the developed instrument through several psychometric aspects, including item difficult, validity, reliability, and potential item bias. Item difficultt reflects a proportional distribution of difficult levels, ranging from items that are easiest to those that are most challenging for students. Validity testing was carried out using three fit indicators within the Rasch model, namely outfit mean square (MNSQ), outfit z-standard (ZSTD), and point measure correlation (PTM Corr) [15]. Instrument reliability indicates the extent to which the measurement tool produces stable and consistent results when administered repeatedly to the same subjects, either across different raters or at different points in time [16]. In addition, Differential Item Functioning (DIF) analysis was employed as a procedure to identify potential item bias, whereby an item may provide a relative advantage or disadvantage to specific groups of respondents compared to others [17].

## 3. RESULTS AND DISCUSSION

The science literacy test instrument that has been empirically tested on 26 students with a total of 20 questions was then analyzed using the Rasch model to test the quality of the instrument in more depth with the help of the Ministep device. Ministep software is a Rasch model based analytical tool designed to process and interpret scores obtained from assessment instruments [18]. According to [19] one of the advantages of the Rasch Model is its ability to provide precise and reliable results in instrument testing. This is because the model uses a probabilistic approach, so that it can recognize the characteristics of the measured object more clearly and consistently. The results of the analysis of instruments based on the model are described in the following section.

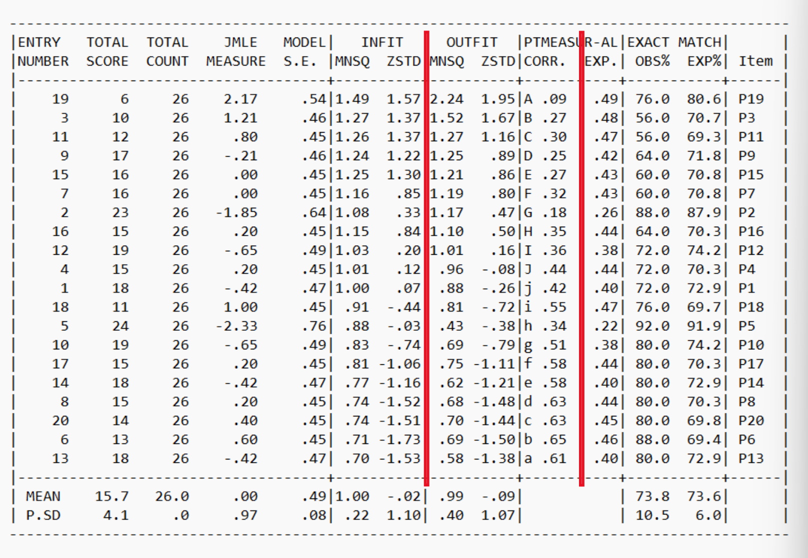
### 3.1. Validity Analysis

Validity testing is a systematic process conducted to evaluate whether an instrument can be considered valid or invalid (Agustin et al., 2025). In this study, validity was examined using three fit indicators within the Rasch model, namely Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (Corr). The criteria for item acceptability were based on the guidelines proposed by [20], as presented in Table 2.

Table 2. Eligibility Ctiteria Item

| Information | Conditions |
|---|---|
| Suitability of each question item | |
| Outfit Mean Square (MNSQ) | 0.50 < x < 1.50 |
| Outfit Z Standard (ZSTD) | -2.0 < x < +2.0 |
| Point Measure Correlation (Corr) | 0.40 < x < 0.85 |

Table 2 presents three criteria used to determine the validity of each instrument item, namely the Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation, along with the acceptable value ranges for item eligibility. Outfit MNSQ values that exceed the acceptable range (0.5–1.5) indicate the presence of misfit or inconsistency in response patterns. Elevated ZSTD values further suggest deviations from the expected model. In addition, Pt Measure Corr values below 0.4 indicate that the items have a weak correlation with respondents' abilities, rendering them less effective in representing the measured construct [21]. Instrument validity was analyzed using Ministep software within the Rasch modeling framework. Among the generated outputs, item fit statistics were examined to evaluate the suitability of each item in measuring the intended construct. The item fit information, obtained from the "Item Fit Order" output, displays items arranged according to their conformity with the Rasch model. These results are summarized in figure 1.



```
------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE   MODEL|   INFIT      OUTFIT   |PTMEASUR-AL|EXACT  MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|-------------------------------------+----------------------+-----------+-----------+------|
|    19      6     26    2.17     .54|1.49  1.57|2.24  1.95|A .09   .49| 76.0  80.6| P19  |
|     3     10     26    1.21     .46|1.27  1.37|1.52  1.67|B .27   .48| 56.0  70.7| P3   |
|    11     12     26     .80     .45|1.26  1.37|1.27  1.16|C .30   .47| 56.0  69.3| P11  |
|     9     17     26    -.21     .46|1.24  1.22|1.25   .89|D .25   .42| 64.0  71.8| P9   |
|    15     16     26     .00     .45|1.25  1.30|1.21   .86|E .27   .43| 60.0  70.8| P15  |
|     7     16     26     .00     .45|1.16   .85|1.19   .80|F .32   .43| 60.0  70.8| P7   |
|     2     23     26   -1.85     .64|1.08   .33|1.17   .47|G .18   .26| 88.0  87.9| P2   |
|    16     15     26     .20     .45|1.15   .84|1.10   .50|H .35   .44| 64.0  70.3| P16  |
|    12     19     26    -.65     .49|1.03   .20|1.01   .16|I .36   .38| 72.0  74.2| P12  |
|     4     15     26     .20     .45|1.01   .12| .96  -.08|J .44   .44| 72.0  70.3| P4   |
|     1     18     26    -.42     .47|1.00   .07| .88  -.26|j .42   .40| 72.0  72.9| P1   |
|    18     11     26    1.00     .45| .91  -.44| .81  -.72|i .55   .47| 76.0  69.7| P18  |
|     5     24     26   -2.33     .76| .88  -.03| .43  -.38|h .34   .22| 92.0  91.9| P5   |
|    10     19     26    -.65     .49| .83  -.74| .69  -.79|g .51   .38| 80.0  74.2| P10  |
|    17     15     26     .20     .45| .81 -1.06| .75 -1.11|f .58   .44| 80.0  70.3| P17  |
|    14     18     26    -.42     .47| .77 -1.16| .62 -1.21|e .58   .40| 80.0  72.9| P14  |
|     8     15     26     .20     .45| .74 -1.52| .68 -1.48|d .63   .44| 80.0  70.3| P8   |
|    20     14     26     .40     .45| .74 -1.51| .70 -1.44|c .63   .45| 80.0  69.8| P20  |
|     6     13     26     .60     .45| .71 -1.73| .69 -1.50|b .65   .46| 88.0  69.4| P6   |
|    13     18     26    -.42     .47| .70 -1.53| .58 -1.38|a .61   .40| 80.0  72.9| P13  |
|-------------------------------------+----------------------+-----------+-----------+------|
| MEAN    15.7   26.0     .00     .49|1.00  -.02| .99  -.09|           | 73.8  73.6|      |
| P.SD     4.1     .0     .97     .08| .22  1.10| .40  1.07|           | 10.5   6.0|      |
------------------------------------------------------------------------------------------
```

Figure 1. Validity of Content on 20 Question Items

Based on Figure 1, the results of the item validity analysis using the Rasch model indicate variability in item fit across the test items. The analysis reports three key fit statistics for each item, namely Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (PTM Corr). A more detailed interpretation of the validity results for each item is systematically presented in Table 3. This presentation is intended to provide a clearer understanding of the extent to which each item satisfies the established validity criteria based on the three Rasch fit indicators (MNSQ, ZSTD, and PTM Corr).

Table 3. Interpretation of Content Validity in 20 Question Items

| Item | Outfit MNSQ | Outfit ZSTD | Ptm Corr. | Conclusion |
|---|---|---|---|---|
| 1 | 0.88 | -0.26 | 0.42 | Question items can be used |
| 2 | 1.17 | 0.47 | 0.18 | Question items can be used |
| 3 | 1.52 | 1.67 | 0.27 | Question items can be used |
| 4 | 0.96 | -0.08 | 0.44 | Question items can be used |
| 5 | 0.43 | -0.38 | 0.34 | Question items can be used |
| 6 | 0.69 | -1.50 | 0.65 | Question items can be used |
| 7 | 1.19 | 0.80 | 0.32 | Question items can be used |
| 8 | 0.68 | -1.48 | 0.63 | Question items can be used |
| 9 | 1.25 | 0.89 | 0.25 | Question items can be used |
| 10 | 0.69 | -0.74 | 0.69 | Question items can be used |
| 11 | 1.27 | 1.16 | 0.30 | Question items can be used |
| 12 | 1.01 | 0.16 | 0.36 | Question items can be used |
| 13 | 0.58 | -1.38 | 0.61 | Question items can be used |
| 14 | 0.62 | -1.16 | 0.62 | Question items can be used |
| 15 | 1.21 | 0.86 | 0.27 | Question items can be used |
| 16 | 1.10 | 0.50 | 0.35 | Question items can be used |
| 17 | 0.75 | -1.11 | 0.58 | Question items can be used |
| 18 | 0.81 | -0.72 | 0.55 | Question items can be used |
| 19 | 2.24 | 1.95 | 0.09 | Question items can be used |
| 20 | 0.70 | -1.44 | 0.63 | Question items can be used |

Based on the analysis presented in Table 3, the majority of items met the validity criteria for Outfit Mean Square (MNSQ) and Outfit Z-Standard (ZSTD), indicating that these items functioned consistently with the Rasch model and were capable of measuring students scientific literacy. However, for several items, the Point Measure Correlation (PTM Corr) values fell below the recommended threshold. Nevertheless, according to [20], items that satisfy at least one of the Rasch fit criteria may still be considered acceptable, whereas items that fail to meet all criteria should be revised or removed. Further analysis revealed that Item 19 exhibited a relatively low Point Measure Correlation (PTM Corr) value compared to other items. As noted by [22], low Point Measure Correlation (PTM Corr) values indicate that an item may not adequately measure the same construct as the remaining items. This condition is likely associated with higher cognitive demands or a problem context that is less familiar to elementary school students. Despite this limitation, Item 19 was retained at the limited trial stage because it still met the overall Rasch model fit criteria. Nonetheless, this item is recommended for revision or close monitoring in future large scale trials. In contrast, Item 5 demonstrated an Outfit Mean Square (MNSQ) value slightly outside the ideal range (0.43). Values below 0.7 are considered indicative of overfitting, suggesting that the response pattern is overly predictable or that the item may be too easy for the target population [23].

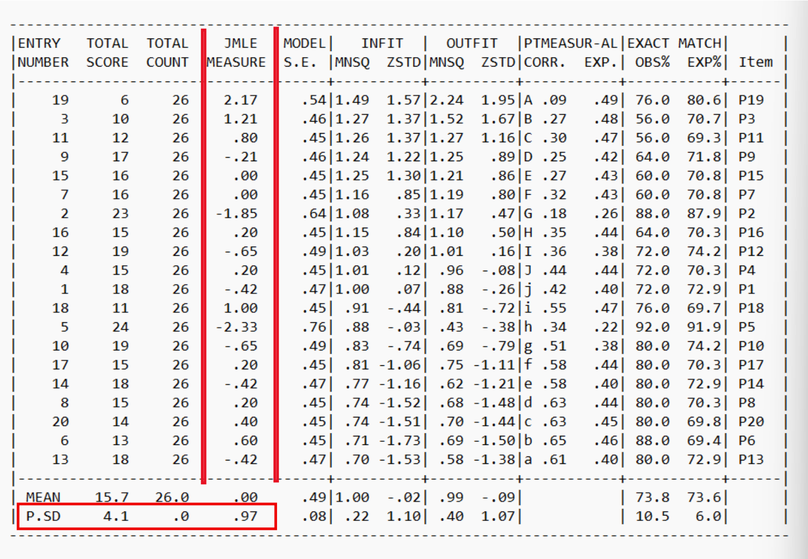
### 3.2. Difficult Level

Item difficulty analysis is the process of examining test items based on their level of ease or difficulty for students, enabling the classification of items into easy, moderate, and difficult categories, as determined by the proportion of students who answer the items correctly rather than by the subjective perceptions of the test developer [24]. The classification of item difficult into four categories follows the criteria proposed by [20], as summarized in Table 4.

Table 4. Interpretation of the difficult level of the question based on the logit value

| Logit Value | Interpretation |
|---|---|
| +1 SD < logit value | Very difficult |
| 0.00 < logit value <+1 SD | Difficult |
| -1 SD < logit value <0.00 | Easy |
| Logit value <-1 SD | It's Very Easy |

Table 4 presents the interpretation of item difficult levels based on logit values obtained from the Rasch model analysis**.** The classification of item difficulty is determined by the position of each item's logit value relative to the mean value (0.00) and the standard deviation (SD). This classification is used to describe the distribution of item difficulty levels and to ensure that the instrument includes an appropriate range of difficult for comprehensively measuring students' abilities. The results of the item difficult analysis were obtained through data processing using Ministep software, which is part of the Rasch model analysis system. This analysis involved examining the output tables generated by the software, with particular attention to Column 15 (Item Measure). The item difficulty measures are subsequently presented in Figure 2.



```
----------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE   MODEL|   INFIT   |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|-----------------------------------+----------+----------+-----------+-----------+------|
|    19      6     26    2.17     .54|1.49  1.57|2.24  1.95|A .09   .49| 76.0  80.6| P19  |
|     3     10     26    1.21     .46|1.27  1.37|1.52  1.67|B .27   .48| 56.0  70.7| P3   |
|    11     12     26     .80     .45|1.26  1.37|1.27  1.16|C .30   .47| 56.0  69.3| P11  |
|     9     17     26    -.21     .46|1.24  1.22|1.25   .89|D .25   .42| 64.0  71.8| P9   |
|    15     16     26     .00     .45|1.25  1.30|1.21   .86|E .27   .43| 60.0  70.8| P15  |
|     7     16     26     .00     .45|1.16   .85|1.19   .80|F .32   .43| 60.0  70.8| P7   |
|     2     23     26   -1.85     .64|1.08   .33|1.17   .47|G .18   .26| 88.0  87.9| P2   |
|    16     15     26     .20     .45|1.15   .84|1.10   .50|H .35   .44| 64.0  70.3| P16  |
|    12     19     26    -.65     .49|1.03   .20|1.01   .16|I .36   .38| 72.0  74.2| P12  |
|     4     15     26     .20     .45|1.01   .12| .96  -.08|J .44   .44| 72.0  70.3| P4   |
|     1     18     26    -.42     .47|1.00   .07| .88  -.26|j .42   .40| 72.0  72.9| P1   |
|    18     11     26    1.00     .45| .91  -.44| .81  -.72|i .55   .47| 76.0  69.7| P18  |
|     5     24     26   -2.33     .76| .88  -.03| .43  -.38|h .34   .22| 92.0  91.9| P5   |
|    10     19     26    -.65     .49| .83  -.74| .69  -.79|g .51   .38| 80.0  74.2| P10  |
|    17     15     26     .20     .45| .81 -1.06| .75 -1.11|f .58   .44| 80.0  70.3| P17  |
|    14     18     26    -.42     .47| .77 -1.16| .62 -1.21|e .58   .40| 80.0  72.9| P14  |
|     8     15     26     .20     .45| .74 -1.52| .68 -1.48|d .63   .44| 80.0  70.3| P8   |
|    20     14     26     .40     .45| .74 -1.51| .70 -1.44|c .63   .45| 80.0  69.8| P20  |
|     6     13     26     .60     .45| .71 -1.73| .69 -1.50|b .65   .46| 88.0  69.4| P6   |
|    13     18     26    -.42     .47| .70 -1.53| .58 -1.38|a .61   .40| 80.0  72.9| P13  |
|-----------------------------------+----------+----------+-----------+-----------+------|
| MEAN    15.7   26.0     .00     .49|1.00  -.02| .99  -.09|           | 73.8  73.6|      |
| P.SD     4.1     .0     .97     .08| .22  1.10| .40  1.07|           | 10.5   6.0|      |
----------------------------------------------------------------------------------------------
```

Figure 2. Difficulty level 20 questions

Based on Figure 2, item difficulty parameters are represented by the JMLE Measure values. According to [25] higher JMLE Measure scores indicate greater item difficulty, whereas lower values correspond to easier items. The item difficulty parameter represents the position of an item on the ability scale, defined as the point at which students have a probability of 0.50 of responding correctly. A higher item difficulty parameter indicates that a higher level of student ability is required to achieve a 50% probability of a correct response [26]. Based on the data processing results presented in Figure 3, the item difficulty levels were subsequently classified to facilitate the interpretation of the analysis results. This classification aimed to group each test item according to its level of difficult based on the item measure values obtained from the Rasch model analysis using Ministep software. Through this classification, the distribution of item difficult across four categories very easy, easy, difficult, and very difficult can be identified. The results of this classification are presented in Table 5, which illustrates the distribution of items within each difficult category.

Table 5. Classification of Difficulty Levels of Question Items

| Item | Item Difficulty Level |
|---|---|
| 19 | Very Difficult |
| 3, 4, 6, 8, 11, 16, 17, 18, 20 | Difficult |
| 1, 2, 7, 9, 10, 12, 13, 14, 15 | Easy |
| 5 | Very Easy |

The distribution of item difficult indicates that most items fall within the moderate category, with a relatively balanced proportion of easy and difficult items. This pattern suggests that the instrument was designed in alignment with the ability level of sixth grade elementary school students, such that it is neither overly easy nor excessively challenging. A balanced distribution of item difficult is essential to ensure that the instrument can effectively differentiate students levels of scientific literacy. This finding is consistent with the view of [27], who emphasize that a high quality test instrument should include items with varied difficult levels namely easy, moderate, and difficult to comprehensively assess students abilities.

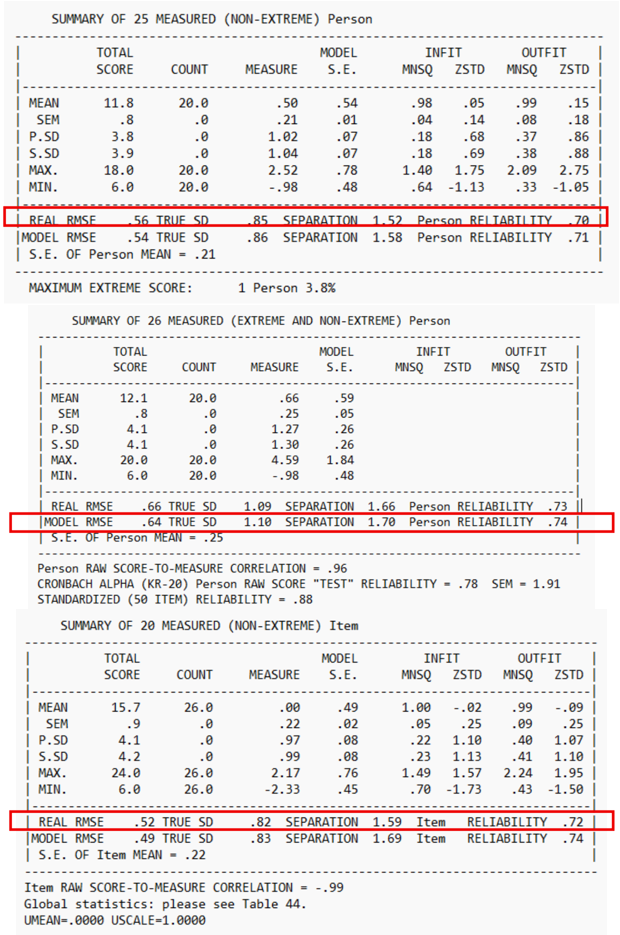
### 3.3. Instrument Reliability

Reliability is an indicator that reflects the degree of trustworthiness of a measurement instrument, referring to the extent to which the instrument is capable of producing consistent and stable data when used repeatedly to measure the same object or variable [28]. The reliability criteria and interpretation proposed by [20] are presented in Table 6.

Table 6. Interpretation of Person Reliability, Item Reliability, and Cronbach Alpha

| Value | Interpretation |
|---|---|
| <0.5 | Bad |
| 0,5 –0,6 | Weak |
| 0,6 –0,7 | Adequate |
| 0,7 –0,8 | Good |
| >0.8 | Very good |

Table 6 presents the guidelines for interpreting instrument reliability based on specific value ranges. Overall instrument reliability is represented by Cronbach's alpha, the consistency of students' responses is indicated by person reliability, and the quality of individual items is reflected by item reliability [29]. The results of the instrument reliability analysis using the Rasch model with the Ministep software are presented in Figure 4.



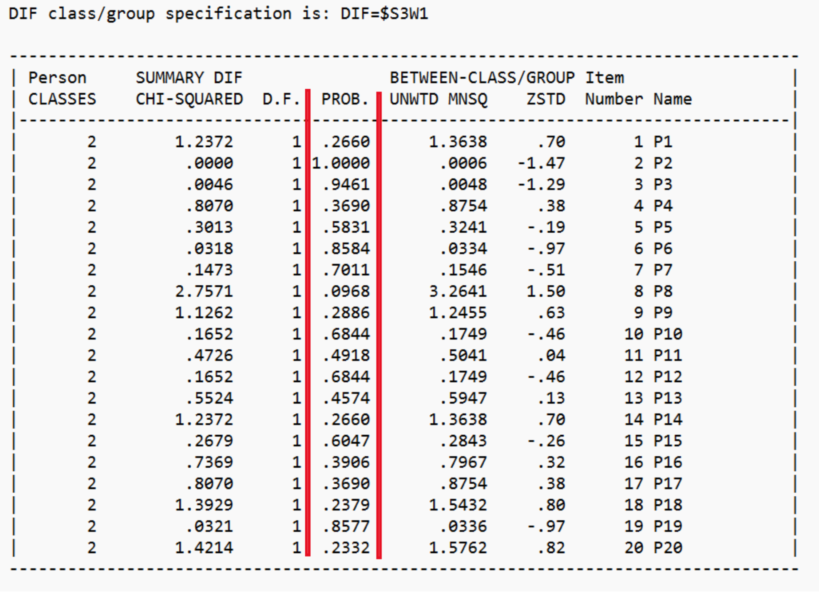
```
   SUMMARY OF 25 MEASURED (NON-EXTREME) Person

-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      11.8      20.0         .50     .54       .98    .05    .99    .15 |
|  SEM        .8        .0         .21     .01       .04    .14    .08    .18 |
| P.SD       3.8        .0        1.02     .07       .18    .68    .37    .86 |
| S.SD       3.9        .0        1.04     .07       .18    .69    .38    .88 |
| MAX.      18.0      20.0        2.52     .78      1.40   1.75   2.09   2.75 |
| MIN.       6.0      20.0        -.98     .48       .64  -1.13    .33  -1.05 |
|-----------------------------------------------------------------------------|
| REAL RMSE     .56 TRUE SD      .85  SEPARATION  1.52  Person RELIABILITY  .70 |
|MODEL RMSE     .54 TRUE SD      .86  SEPARATION  1.58  Person RELIABILITY  .71 |
| S.E. OF Person MEAN = .21                                                   |
-------------------------------------------------------------------------------
 MAXIMUM EXTREME SCORE:      1 Person 3.8%

     SUMMARY OF 26 MEASURED (EXTREME AND NON-EXTREME) Person

-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      12.1      20.0         .66     .59                                |
|  SEM        .8        .0         .25     .05                                |
| P.SD       4.1        .0        1.27     .26                                |
| S.SD       4.1        .0        1.30     .26                                |
| MAX.      20.0      20.0        4.59    1.84                                |
| MIN.       6.0      20.0        -.98     .48                                |
|-----------------------------------------------------------------------------|
| REAL RMSE     .66 TRUE SD     1.09  SEPARATION  1.66  Person RELIABILITY  .73 ||
|MODEL RMSE     .64 TRUE SD     1.10  SEPARATION  1.70  Person RELIABILITY  .74 |
| S.E. OF Person MEAN = .25                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .78  SEM = 1.91
STANDARDIZED (50 ITEM) RELIABILITY = .88

    SUMMARY OF 20 MEASURED (NON-EXTREME) Item

-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      15.7      26.0         .00     .49      1.00   -.02    .99   -.09 |
|  SEM        .9        .0         .22     .02       .05    .25    .09    .25 |
| P.SD       4.1        .0         .97     .08       .22   1.10    .40   1.07 |
| S.SD       4.2        .0         .99     .08       .23   1.13    .41   1.10 |
| MAX.      24.0      26.0        2.17     .76      1.49   1.57   2.24   1.95 |
| MIN.       6.0      26.0       -2.33     .45       .70  -1.73    .43  -1.50 |
|-----------------------------------------------------------------------------|
| REAL RMSE     .52 TRUE SD      .82  SEPARATION  1.59  Item   RELIABILITY  .72 |
|MODEL RMSE     .49 TRUE SD      .83  SEPARATION  1.69  Item   RELIABILITY  .74 |
| S.E. OF Item MEAN = .22                                                     |
-------------------------------------------------------------------------------
Item RAW SCORE-TO-MEASURE CORRELATION = -.99
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000
```

Figure 4. Reliability in 20 question items

Based on Figure 4, statistical information is obtained, including the mean score, standard deviation, maximum and minimum values, as well as the reliability index and separation levels. The reliability criteria in this analysis are determined using Cronbach's Alpha, which represents internal consistency and reflects the overall interaction between respondents and instrument items. The reliability analysis results indicate that the person reliability value was 0.70, while the item reliability value was 0.72. Based on the reliability interpretation criteria presented in Table 6, reliability values ranging from 0.6 to 0.7 are classified as *adequate*, whereas values between 0.7 and 0.8 are considered *good*. Accordingly, these findings suggest that the consistency of students responses to the test items falls within the adequate category, while the stability of the instrument in measuring the assessed construct is categorized as good. In addition, the Cronbach's Alpha coefficient of 0.78 further supports these results, as it also falls within the good category. Overall, these findings indicate that the developed test items satisfy the reliability requirements. From a practical perspective, this suggests that the instrument can be utilized by teachers to obtain an initial overview of students abilities and to monitor the development of scientific literacy related to the topic of global warming.

### 3.4 Differential Item Function (DIF)

Differential Item Functioning (DIF) is an analytical approach aimed at identifying systematic differences in item characteristics when test items are responded to by groups of participants with differing backgrounds or characteristics [30]. In this study, the DIF analysis focused on detecting potential item bias related to gender differences among students. The results of this analysis are presented in detail in Figure 5.



```
DIF class/group specification is: DIF=$S3W1

---------------------------------------------------------------------------------
| Person     SUMMARY DIF                    BETWEEN-CLASS/GROUP Item             |
| CLASSES    CHI-SQUARED  D.F.  PROB.  UNWTD MNSQ     ZSTD  Number Name          |
|--------------------------------------------------------------------------------|
|        2       1.2372      1  .2660      1.3638      .70       1 P1            |
|        2        .0000      1 1.0000       .0006   -1.47       2 P2            |
|        2        .0046      1  .9461       .0048   -1.29       3 P3            |
|        2        .8070      1  .3690       .8754      .38       4 P4            |
|        2        .3013      1  .5831       .3241     -.19       5 P5            |
|        2        .0318      1  .8584       .0334     -.97       6 P6            |
|        2        .1473      1  .7011       .1546     -.51       7 P7            |
|        2       2.7571      1  .0968      3.2641     1.50       8 P8            |
|        2       1.1262      1  .2886      1.2455      .63       9 P9            |
|        2        .1652      1  .6844       .1749     -.46      10 P10           |
|        2        .4726      1  .4918       .5041      .04      11 P11           |
|        2        .1652      1  .6844       .1749     -.46      12 P12           |
|        2        .5524      1  .4574       .5947      .13      13 P13           |
|        2       1.2372      1  .2660      1.3638      .70      14 P14           |
|        2        .2679      1  .6047       .2843     -.26      15 P15           |
|        2        .7369      1  .3906       .7967      .32      16 P16           |
|        2        .8070      1  .3690       .8754      .38      17 P17           |
|        2       1.3929      1  .2379      1.5432      .80      18 P18           |
|        2        .0321      1  .8577       .0336     -.97      19 P19           |
|        2       1.4214      1  .2332      1.5762      .82      20 P20           |
---------------------------------------------------------------------------------
```

Figure 5. Differential Item Function (DIF) Analysis

The criterion used in this analysis is the probability value (Prob.) for each test item, as presented in Figure 5. According to Prayoga et al [20], if the Prob. value is less than 0.05, the item is considered to exhibit Differential Item Functioning (DIF), indicating that differences in the probability of responding correctly are not solely attributable to differences in students' ability but are influenced by group characteristics, such as gender. Conversely, if the Prob. value is greater than or equal to 0.05, the item can be regarded as not exhibiting DIF or as being free from bias.

Based on the results presented in the figure, all test items exhibited Prob. values greater than 0.05, indicating that the developed instrument does not show any evidence of gender related bias. The absence of gender bias in items related to the topic of global warming has important pedagogical implications. According to Setiawan et al [31], global warming is a multidimensional phenomenon that integrates scientific concepts, environmental issues, and everyday life contexts. As such, global warming content is inherently contextual and applied, and is less influenced by strong gender based social constructs [32].

## 4. CONCLUSION

The context based scientific literacy skills test on the topic of global warming developed in this study demonstrates preliminary evidence of being a valid, reliable, and feasible instrument based on Rasch model analysis. The results indicate that the majority of items met the validity criteria as reflected in the Outfit Mean Square (MNSQ), Outfit Z-Standardized (ZSTD), and Point Measure Correlation (PTM Corr) indices. The distribution of item difficulty was proportionally spread across categories ranging from very easy to very difficult, suggesting that the instrument was designed to align with the ability range of Grade VI elementary school students. Furthermore, the person reliability value of 0.70, item reliability of 0.72, and Cronbach's alpha of 0.78 indicate adequate to good internal consistency. The Differential Item Functioning (DIF) analysis also revealed that all items were free from gender bias, indicating that the instrument meets the principle of measurement fairness. Nevertheless, as this study constitutes a pilot study involving a limited number of respondents from a single school context, the findings should be interpreted as preliminary evidence rather than as a final validation of the instrument. Therefore, future research is strongly recommended to conduct large-scale empirical testing with more diverse and representative samples and to apply the instrument across different school contexts in order to examine the stability of item parameters, strengthen construct validity, and enhance the generalizability of the findings.

## ACKNOWLEDGEMENTS

The author would like to thank all parties who have contributed to the completion of this research. Particular thanks are extended to the Principal for the permission and support provided.

## AUTHOR CONTRIBUTIONS

Conceptualization, Methodology, Formal Analysis, Investigation, Resources, Data Curation, Writing-Original Draft Preparation, & Visualization, MNF; Writing – Review & Editing, MNF, RPRH and ES; Supervision & Project Administration, MNF, RPRH, ES.

## CONFLICTS OF INTEREST

The author(s) declare no conflict of interest.

## USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

The authors declare that no artificial intelligence (AI) tools were used in the generation, analysis, or writing of this manuscript. All aspects of the research, including data collection, interpretation, and manuscript preparation, were carried out entirely by the authors without the assistance of AI-based technologies.

## REFERENCES

[1] F. Kasse and I. R. W. Atmojo, “Analisis kecakapan abad 21 melalui literasi sains pada siswa sekolah dasar [Analysis of 21st Century Skills through Scientific Literacy in Elementary School Students],” *J. Educ. Dev. Pendidik. Tapanuli Selatan*, vol. 10, no. 1, pp. 124–128, 2022, doi: 10.37081/ed.v10i1.3322.

[2] I. S. Budiarti and Tanta, “Analysis on students’ scientific literacy of newton’s law and motion system in living things,” *J. Pendidik. Sains Indones.*, vol. 9, no. 1, pp. 36–51, 2021, doi: 10.24815/jpsi.v9i1.18470.

[3] I. Irsan, “Implemensi literasi sains dalam pembelajaran IPA di sekolah dasar [Implementation of scientific literacy in science learning at elementary school],” *J. Basicedu*, vol. 5, no. 6, pp. 5631–5639, 2020, doi: 10.31004/basicedu.v5i6.1682.

[4] C. Z. L. Parisu, L. Sisi, and A. Juwairiyah, “Pengembangan literasi sains pada siswa sekolah dasar melalui pembelajaran IPA [Development of scientific literacy in elementary school students through science learning],” *J. Pendidik. Multidisiplin*, vol. 1, no. 1, pp. 11–19, 2025, doi: 10.54297/jpmd.v1i1.880.

[5] C. Z. L. Parisu, E. E. Saputra, and Larisisi, “Integrasi pendidikan karakter dalam pembelajaran ipa di sekolah dasar [Integration of character education in science learning at elementary school],” *Pendas J. Ilm. Pendidik. Dasar*, vol. 5, no. 1, pp. 864–872, 2025, doi: 10.23969/jp.v8i1.7488.

[6] S. Ayub, J. Rokhmat, A. Ramdani, and A. Hakim, “Karakteristik soal literasi sains programme for international student assesment (PISA) tahun 2015 [Characteristics of science literacy items in PISA 2015],” *J. Ilm. Profesi Pendidik.*, vol. 7, no. 4b, pp. 2623–2629, 2022, doi: 10.29303/jipp.v7i4b.1039.

[7] T. Tanti, W. Utami, D. Deliza, and M. Jahanifar, “Investigation in vocation high school for attitude and motivation students in learning physics subject”, *Journal Evaluation in Education (JEE),* vol. 6, no. 2, pp. 479-490, 2025, doi: 10.37251/jee.v6i2.1452.

[8] H. Fuadi, A. Z. Robbia, J. Jamaluddin, and A. W. Jufri, “Analisis faktor penyebab rendahnya kemampuan literasi sains peserta didik [Analysis of the factors contributing to students’ low science literacy skills],” *J. Ilm. Profesi Pendidik.*, vol. 5, no. 2, pp. 108–116, 2020, doi: 10.29303/jipp.v5i2.122.

[9] F. Avvisati and R. Ilizaliturri, “PISA 2022 Results: Factsheets – Indonesia,” 2023.

[10] Hasnawati, M. Syazali, and G. P. Putra, “Pengembangan asesmen literasi sains berbasis PISA untuk siswa sekolah dasar [Development of PISA-Based scientific literacy assessment for elementary school students],” *J. Ilm. Pendidik. Dasar*, vol. 5, no. 2, pp. 240–250, 2023, doi: 10.37216/badaa.v5i1.1213.

[11] I. Magdalena, M. Hifziyah, V. N. Aeni, and R. P. Rahayu, “Pengembangan instrumen tes siswa tingkat sekolah dasar Kabupaten Tangerang [Development of a test instrument for elementary school students in Tangerang Regency],” *J. Pendidik. dan Ilmu Sos.*, vol. 2, no. 2, pp. 227–237, 2020, doi: 10.36088/nusantara.v3i2.1244.

[12] L. Zhang, X. Liu, and H. Feng, “Development and validation of an instrument for assessing scientific literacy from junior to senior high school,” *Discip. Interdiscip. Sci. Educ. Res.*, vol. 5, no. 21, pp. 2–15, 2023, doi: 10.1186/s43031-023-00093-2.

[13] H. B. Atta, Vlorensius, A. Irianto, and Ikhsanudin, “Developing an instrument for students scientific literacy Developing an instrument for students scientific literacy,” *J. Phys. Conf. Ser.*, 2020, doi: 10.1088/1742-6596/1422/1/012019.

[14] N. N. Faidah, M. Listiawati, and I. M. Yamin, “Pengaruh penggunaan media pembelajaran liveworksheets dalam meningkatkan hasil belajar kognitif siswa pada materi pemanasan global [The effect of using liveworksheets as an instructional media on improving students’ cognitive learning outcomes in global warming material],” *J. Kiprah Pendidik.*, vol. 1, no. 2, pp. 194–208, 2023, doi: 10.33578/kpd.v2i2.182.

[15] M. Yusup, “Using Rasch model for the development and validation of energy literacy assessment instrument for prospective physics teachers Using Rasch model for the development and validation of energy literacy assessment instrument for prospective physics teachers,” *J. Phys. Conf. Ser.*, pp. 1–10, 2021, doi: 10.1088/1742-6596/1876/1/012056.

[16] E. Nazharita *et al.*, “Analisis butir soal tes literasi numerasi domain geometri siswa SMP menggunakan rasch model berbantuan ministep [Item analysis of numeracy literacy test in geometry domain for junior high school students using the rasch model assisted by ministep],” *J. Pendidik. Mat.*, vol. 10, no. 3, pp. 989–1000, 2025, doi:

10.30605/pedagogy.v10i3.6785.

[17] Y. Helsa and D. Fitria, *Pengantar Statistik Untuk Mahasiswa Pendidikan Guru Sekolah Dasar dan Umum Jilid 2 [Introduction to Statistics for Elementary School Teacher Education and General Students, Volume 2].* Yogyakarta: Deepublish Digital, 2024.

[18] W. Lestari, I. Wigati, M. I. Sholeh, and D. Pramita, “Instrumen literasi digital guru menggunakan model rasch [Teachers’ digital literacy instrument using the rasch model],” *Orbital J. Pendidik. Kim.*, vol. 6, no. 2, pp. 104–113, 2022, doi: 10.19109/ojpk.v6i2.15019.

[19] N. Monigir, Y. W. Tumbol, and N. V. Monolimay, “Use of The Rasch Model for Analysis of Test Instruments in Class V Science Subjects State 38 Primary School Manado,” *Int. J. Inf. Technol. Educ.*, vol. 3, no. 2, pp. 1–7, 2024, doi: 10.62711/ijite.v3i2.171.

[20] K. P. Prayoga, D. Suryana, M. Supriatna, and N. Budiman, “Penggunaan Rasch Model Untuk Menganalisis Konstruk Instrumen Kontrol Diri Pada Siswa Sekolah Menengah [The Use of the Rasch Model to Analyze the Construct of a Self-Control Instrument for Secondary School Students],” *G-Couns J. Bimbing. dan Konseling*, vol. 9, no. 1, pp. 367–381, 2024, doi: 10.31316/gcouns.v9i1.4459.

[21] B. Sumintono and W. Widhiarso, *Aplikasi Permodelan Rasch Pada Assesment Pendidikan [Application of Rasch Modeling in Educational Assessment]*. Cimahi: Penerbit Trim Komunikata, 2015.

[22] T. Tanti, A. Astalini, D. A. Kurniawan, D. Darmaji, T. O. Puspitasari, and I. Wardhana, “Attitude for physics: The condition of high school students,” *Jurnal Pendidikan Fisika Indonesia,* vol. 17, no. 2, pp. 126-132, 2021, doi: 10.15294/jpfi.v17i2.18919.

[23] H. Noperi, V. Wiliyanti, and F. A. Yanti, “Validasi Instrumen Literasi Sains dengan Model Rasch untuk Socio-Scientific Issues [Validation of a Scientific Literacy Instrument Using the Rasch Model for Socio-Scientific Issues],” *Nat. J. Ilm. Pendidik. IPA*, vol. 11, no. 2, pp. 63–75, 2024, doi: 10.30738/natural.v11i2.18891.

[24] R. E. Tornabene, E. Lavington, and R. H. Nehm, “Testing validity inferences for Genetic Drift Inventory scores using Rasch modeling and item order analyses,” *Evol. Educ. Outreach*, vol. 11, no. 6, pp. 2–16, 2018, doi: 10.1186/s12052-018-0082-x.

[25] T. Tanti, A. Astalini, D. Darmaji, D. A. Kurniawan, and R. Fitriani, “Student perception review from gender: Electronic moduls of mathematical physics,” *JPI (Jurnal Pendidikan Indonesia)*, vol. 11, no. 1, pp. 125-132, 2022, doi: 10.23887/jpiundiksha.v11i1.35107.

[26] R. Athiyyah, S. Feranie, and T. R. Ramalis, “Development of scientific process-creative skills (sp-cs) test on light wave concept: Content validity and rasch model analysis,” *Edusains*, vol. 14, no. 2, pp. 111–125, 2022, doi: 10.15408/es.v13i2.28025.

[27] H. Hadiyanti, P. Susongko, and Munadi, “Pengembangan instrumen higher order thinking skill mata pelajaran matematika dengan rasch model [Development of a higher order thinking skills instrument in mathematics using the rasch model],” *J. Educ. Res.*, vol. 5, no. 1, pp. 399–407, 2024, doi: https://doi.org/10.37985/jer.v5i1.765.

[28] R. A. Setiyawan and P. S. Wijayanti, “Analisis kualitas instrumen untuk mengukur kemampuan pemecahan masalah siswa selama pembelajaran daring di masa pandemi [Analysis of Instrument Quality to Measure Students’ Problem-Solving Skills during Online Learning in the Pandemic Period],” *Lebesgue JurnalIlmiahPendidikan Mat. Mat. dan Stat.*, vol. 1, no. 2, pp. 130–139, 2020, doi: https://doi.org/10.46306/lb.v1i2.26.

[29] H. A. Saputri, Zulhijrah, N. J. Larasati, and Shaleh, “Analisis instrumen assesmen: Validitas, reliabilitas, tingkat kesukaran, dan daya beda butir soal [Assessment instrument analysis: Validity, reliability, difficulty level, and discrimination index],” *J. Ilm. PGSD FKIP Univ. Mandiri*, vol. 9, no. 5, pp. 2986–2995, 2023, doi: 10.36989/didaktik.v9i5.2268.

[30] R. I. Octaviana, M. B. Anggara, R. Jamilah, A. Darmana, and R. D. Suyanti, “Analisis item soal kimia SMA menggunakan rasch model “Analisis item soal kimia SMA menggunakan rasch model [Analysis of senior high school chemistry test items using the rasch model],” *ORBITAL J. Pendidik. Kim.*, vol. 6, no. 1, pp. 26–37, 2022, doi: 10.19109/ojpk.v6i1.12248.

[31] M. A. Setiawan, P. Susongko, and M. N. H. Hayati, “Pendeteksian DIF pada perangkat tes objektif penilaian akhir semester ipa dengan mengunakan permodelan rasch [Detection of DIF in science final semester objective test using rasch modeling],” *PSEJ (Pancasakti Sci. Educ. Journal)*, vol. 5, no. 2, pp. 23–29, 2020, doi: 10.24905/psej.v5i2.25.

[32] Harpina, S. A. Darfin, and U. N. Kholifatun, “Science literacy and climate change issues in elementary school science learning as a green education effort,” *J. Humanit. Soc. Sci. Educ.*, vol. 1, no. 2, pp. 55–68, 2025, doi: 10.64690/jhuse.v1i2.34.