14 (2) (2025) 405-416

Unnes Science Education Journal

Accredited Sinta 2

https://journal.unnes.ac.id/journals/usej

Feasibility Analysis of Scientific Reasoning Skills Instruments for Prospective Science Teachers Using The Rasch Model

Yuni Arfiani[1,2]✉, Dwi Yulianti[1], Masturi Masturi[1], Sunyoto Eko Nugroho[1]

DOI: http://dx.doi.org/10.15294/usej.v13i1.25493

[1]Universitas Negeri Semarang, Indonesia
[2]Universitas Pancasakti Tegal, Indonesia

Article Info

Submitted 2025-06-10
Revised 2025-07-05
Accepted 2025-08-01

Keywords

Prospective Science Teacher; Scientific Reasoning Skills; Rasch Model; Reliability; Validity

Abstract

This study aims to analyze the feasibility of a scientific reasoning ability test instrument for prospective science teachers using the Rasch model. The instrument developed refers to Lawson's Classroom Test of Scientific Reasoning (LCTSR) indicators, focusing on scientific reasoning abilities, including variable control, combinatorial, probability, relational, and proportional thinking. The analysis was carried out using the Rasch Model with the WINSTEPS Version 5.7.3.0 application to test the validity, reliability, and characteristics of the test items. The feasibility test was carried out on 40 multiple-choice questions, tested on 110 test samples. The analysis results showed that 25 of the 40 questions were valid and covered all indicators of problem-solving skills. The test instrument also has a high level of internal consistency with a person reliability value of 0.84 and an item reliability of 0.95. Analysis of the level of difficulty of the test items showed a balanced distribution between very difficult, difficult, moderate, and easy questions. No items were detected as biased, indicating that this instrument is fair and can be used without favoring one group of individuals. Overall, the developed problem-solving skills test instrument is valid and reliable for use on prospective science teachers, assisting educators in evaluating students' problem-solving skills, providing constructive feedback, and improving the quality of learning in science.

How to Cite

Arfiani, Y., Yunianti, D., Masturi, M., & Nugroho, S. E. (2025). Feasibility Analysis of Scientific Reasoning Skills Instruments for Prospective Teacher Using The Rasch Model. *Unnes Science Education Journal*, 14(2), 405-416.

✉ Correspondence Author:
E-mail: yuniarfiani@students.unnes.ac.id

p-ISSN 2252-6617
e-ISSN 2502-6232

INTRODUCTION

Scientific reasoning skills in the 21st century are becoming increasingly important in facing the complexities and challenges that exist. This era is marked by rapid technological change, globalization, and various social and environmental problems that require innovative and collaborative solutions (Anjani et al., 2020; Kara & Aslan, 2024; E. Yulianti & Zhafirah, 2020). Related to the complexity of global problems, problems such as climate change, energy crisis, global health, and social inequality require a holistic and integrated problem-solving approach. Solving 21st century problems requires considering various aspects, such as the environment, economy, and society, so good scientific reasoning skills are needed. Technological advances such as artificial intelligence (AI), big data, and the internet of things (IoT) bring new and complex challenges. As part of the higher-order thinking process, scientific reasoning is a highly relevant concept in the discipline of education (Díaz et al., 2023). Good scientific reasoning skills are needed to understand and manage these technologies and overcome problems that may arise from their use. In the context of education, scientific reasoning skills, together with problem-solving skills, are closely related to academic success. Students who have good scientific reasoning will be able to solve problems well, so they tend to achieve higher achievements, especially in subjects such as mathematics, science, and engineering (Nugroho & Waslam, 2020).

Education has become a basic need in this modern era. Without education, people will have difficulty keeping up with developments in science and technology and will have difficulty getting jobs. To realize good education, an appropriate learning process is needed in every educational activity. The learning process includes three aspects, namely learning methods, learning media, and assessment instruments (Hadarah & Tulhikmah, 2019; D. Yulianti et al., 2020) . Based on the current education curriculum in Indonesia, the 2013 Curriculum, one of the learning models that can be used according to the Regulation of the Minister of Education and Culture Number 22 of 2016, is the 2013 Curriculum. Problem-Based Learning Model (PBL).

Scientific reasoning ability measurement instruments play a very important role in education and various other sectors. These instruments allow educators and trainers to evaluate the extent to which a person has mastered these abilities. This is important to understand an individual's strengths and weaknesses so that additional assistance or training can be provided if needed (Ben-Horin et al., 2023; Zulkipli et al., 2020). By systematically measuring students' scientific reasoning abilities, schools and educational institutions can evaluate the effectiveness of their curriculum and teaching methods (Yanto et al., 2019). This information can be used to make necessary adjustments to improve the quality of education. The use of these instruments allows educators to monitor students' academic progress over time. With continuous monitoring, educators can provide constructive feedback and help students to continue to develop.

This study analyzed the feasibility of the scientific reasoning ability test instrument designed for prospective science teachers. The test instrument developed refers to scientific reasoning indicators according to Lawson's Classroom Test of Scientific Reasoning (LCTSR), focusing on scientific reasoning abilities, including variable control, combinatorial thinking, probability thinking, relational thinking, and proportional thinking (Lawson, 1978; Suaidah et al., 2023; Zhou et al., 2021). The feasibility of the developed instrument is reviewed from the validity, reliability, and characteristics of the test items analyzed using the Rasch Model (RM). The Rasch model (RM) analysis, which is an item response theory (IRT) model, was developed by Georg Rasch around 1960. RM analysis accommodates a probability approach in viewing the measurement subject so that RM analysis is not deterministic and can identify the subject being measured more accurately. In addition, Rasch modeling can overcome differences in metrics between items and overcome data interval problems. Therefore, this study was conducted to analyze ICS using RM to provide information on the feasibility of the scientific reasoning test instrument using the Rasch model. Thus, the instrument can be evaluated to obtain the best measuring instrument to determine the scientific reasoning ability for prospective science teachers.

Despite growing awareness of importance of scientific reasoning, various studies have indicated that many prospective science teachers still demonstrate low to moderate levels of scientific reasoning skills (Bao et al., 2022; Jufri et al., 2016). While several instruments have been developed to measure these skills, many of them lack empirical evidence of psychometric feasibility, particularly those validated using modern measurement models such as Rasch Model (Ben-Horin et al., 2023; Maulana et al., 2023). Moreover, previous research often emphasizes application of scientific reasoning in classroom contexts but rarely focuses on assessing robustness and fairness of instruments themselves through rigorous statistical modeling (Firdaus et al., 2025; Thayaseelan et al., 2024; Winari & Masturi, 2023).

This research addresses that gap by not only focusing on measuring scientific reasoning but also by evaluating the instrument's psychometric properties comprehensively using the Rasch Model. The novelty of this study lies in the development and validation of a scientific reasoning instrument specifically tailored for prospective science teachers in Indonesian educational context, analyzed through item response theory. Most prior studies still rely heavily on classical test theory (CTT), which tends to be sample-dependent and lacks diagnostic precision offered by Rasch Model (Jumadi et al., 2023). The use of the Rasch Model provides a more objective and generalizable understanding of item and person fit, as well as the scale's validity and reliability.

The urgency of this research is further underscored by the growing demand for 21st century competencies in teacher education. Prospective science teachers must be equipped not only with content knowledge but also with metacognitive and reasoning abilities to navigate socio-scientific issues and technological advancements. Therefore, a valid and reliable assessment instrument is essential to monitor and improve the development of these competencies in teacher preparation programs (Fütterer et al., 2023). Furthermore, in the wake of Indonesia's Merdeka Belajar curriculum, which emphasizes higher-order thinking skills, it becomes increasingly important to ensure that scientific reasoning can be systematically and fairly assessed using appropriate tools.

METHOD

The focus of this research is related to the development of scientific reasoning test instruments for prospective science teachers. This test instrument was developed from Lawson's Classroom Test of Scientific Reasoning (LCTSR) indicators, focusing on scientific reasoning abilities, including variable control, combinatorial thinking, probability thinking, relational thinking, and proportional thinking. These indicators are described and developed into 40 multiple-choice test items that go through a feasibility testing stage. The research stages are presented in Figure 1.

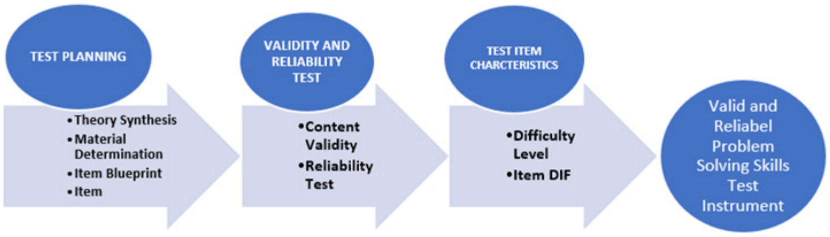


Figure 1. Scientific Reasoning Ability Test Instrument Development Stages

In the initial stage, a literature review was conducted related to indicators of scientific reasoning ability. After finding the appropriate indicators, a blueprint item was created. The indicators of scientific reasoning ability used Lawson's Classroom Test of Scientific Reasoning (LCTSR) indicators, which were then described in Table 1. The description of indicators in Table 1 was used as basis for creating 40 multiple-choice questions that were adjusted to cognitive dimensions and material indicators.

Table 1. Distribution of Question Items Based on Scientific Reasoning Ability Indicators

Indicator	Decription	Item Number	Questions
Control of Variables	Refers to the skills to identify, manipulate, and control independent, dependent, and control variables in a scientific experiment.	1,2,5,6,7,8, 11,17,25,26	10
Combi-natorial Thinking	Reflects an individual's capacity to think about and calculate all possible combinations of various variables or elements available in a situation.	12,13,14,32, 33,34, 16,27	8
Probability Thinking	Relating to understanding, interpreting, and applying concept of chance or probability in explaining scientific phenomena or in making predictions based on data.	3,4,9,10,15, 21,36,37, 39	9
Relational Thinking	refers to the skill of recognizing and explaining the relationship between two or more variables.	18,19,20,29, 30,31	6
Proportional Thinking	Includes understanding and application of concepts of ratio, comparison, and proportion explaining relationship between variables.	22,23,24,28, 35,38,40	7

The test items that have been prepared are then subjected to a feasibility test consisting of a validity test, reliability test, item response test, and question characteristic test. The test was conducted using Rasch Item Fit Model with WINSTEPS Version 5.7.3.0 application. The output data obtained from the WINSTEPS application was then interpreted to determine valid questions and level of instrument reliability. The criteria used to select appropriate or valid questions were carried out by looking at Outfit Mean Square (MNSQ), Z-Standard Outfit (ZSTD), and Outfit Point Value Correlation (PT Measure Corr) values (Boone, 2016). The criteria used are:

1. Outfit Mean Square (MNSQ) received: 0.5 <MNSQ <1.5
2. Z-Standard Outfit (ZSTD) value received: –2.0 <ZSTD <+2.0
3. Outfit Point Value Correlation (Pt Mean Corr) Value: 0.4 <Pt Measure Corr <0.85

Instrument items are said to be valid if they meet at least 2 of the acceptance criteria. If an item is found whose MNSQ and PT Measure Corr values do not meet but whose ZCTD values meet the criteria, then the item is still considered fit and can still be used.

Reliability analysis consists of person reliability and item reliability. The categories of person reliability and item reliability refer to Table 2. (Boone, et.al, 2016).

Table 2. Person reliability and item reliability categories

Item person reliability and item reliability values	Category
<0.67	Weak
0.67-0.80	Enough
0.80-0.90	Good
0.91-0.94	Very Good
>0.94	Excellent

Interpretation of the internal consistency of the test items was carried out based on the Cronbach’s Alpha Coefficient in Table 3. The next stage after the validity and reliability test is the item response test (IRT), and the test of the level of difficulty of the questions and the test of the bias of the questions (item DIF) (Abate et al., 2020). IRT is carried out using 3 parameters, namely unidimensionality, local independence, and parameter invariance, while the test of the level of difficulty of the questions uses Item Measure on the WINSTEPS Version 5.7.3.0 application. Determination of the level of difficulty of the questions is based on the standard deviation (SD) value and the Logit value.

Table 3. The Classification of Cronbach’s Alpha Coefficient

Cronbach's Alpha Coefficient	Interpretation
<0.5	The scale has no internal consistency
0.5-0.6	The internal consistency of the scale is weak
0.6-0.7	The internal consistency of the scale is acceptable
0.7-0.8	The scale has internal consistency
>0.8	The internal consistency of the scale is very high

RESULT AND DISCUSSION

Content Validity

The scientific reasoning test instrument that was prepared consisted of 5 indicators that were adjusted to the material of Water Pollution. The fit items of the content validity test results of the scientific reasoning ability instrument can be seen in Table 2. The results showed that there were 25 out of 40 fit questions, so these items were declared valid and could be used as the right items in measuring scientific reasoning ability.

Table 4 Item Fit Results of Scientific Reasoning Instrument Validity Test

Item	MNSQ	ZSTD	Pt. Measure Corr	Description	Item	MNSQ	ZSTD	Pt. Measure Corr	Description
1	2.46	4.70	-.22	not valid	21	0.96	-0.19	0.50	valid
2	0.57	-1.25	0.44	valid	22	1.25	2.20	0.18	not valid
3	0.75	-2.64	0.58	valid	23	1.59	5.09	-.06	not valid
4	2.68	2.55	.03	not valid	24	0.95	-0.20	0.42	valid
5	0.49	-2.49	0.61	not valid	25	0.84	-1.31	0.57	valid
6	0.43	-2.29	0.61	not valid	26	0.84	-1.61	0.53	valid
7	1.75	3.31	-.15	not valid	27	1.26	1.06	0.32	valid

Item	MNSQ	ZSTD	Pt. Measure Corr	Description	Item	MNSQ	ZSTD	Pt. Measure Corr	Description
8	0.53	-1.40	0.41	valid	21	0.96	-0.19	0.50	valid
9	0.49	-1.03	0.33	valid	22	1.25	2.20	0.18	not valid
10	0.79	-1.26	0.66	valid	23	1.59	5.09	-.06	not valid
11	1.46	4.04	.07	not valid	24	0.95	-0.20	0.42	valid
12	0.40	-0.83	0.28	not valid	25	0.84	-1.31	0.57	valid
13	1.31	1.23	.29	valid	26	0.84	-1.61	0.53	valid
14	0.75	-2.66	0.59	valid	27	1.26	1.06	0.32	valid
15	1.56	2.62	.31	not valid	35	1.00	0.10	0.34	valid
16	0.42	-3.94	0.80	not valid	36	1.10	.90	0.35	valid
17	0.57	-2.85	0.67	valid	37	1.26	2.44	0.19	not valid
18	1.01	.10	0.40	valid	38	1.09	.57	0.45	valid
19	0.68	-2.79	0.61	valid	39	0.76	-2.49	0.59	valid
20	1.30	1.91	.32	valid	40	0.97	-0.25	0.44	valid

Based on Table 4, valid questions are then grouped according to the indicators of scientific reasoning ability. The results are in Table 5.

Table 5. Grouping of Valid Test Items according to Indicators

Indicator	Valid Question Number	Number of Questions
Control of Variable	2,8,17,25,26	5
Combinatorial Thinking	13,14,32,33,27	5
Probability Thinking	3,9,10,21,36, 39	6
Relational Thinking	18,19,20,30,31	5
Proportional Thinking	38,40,24,35	4
Number of Valid Questions		25

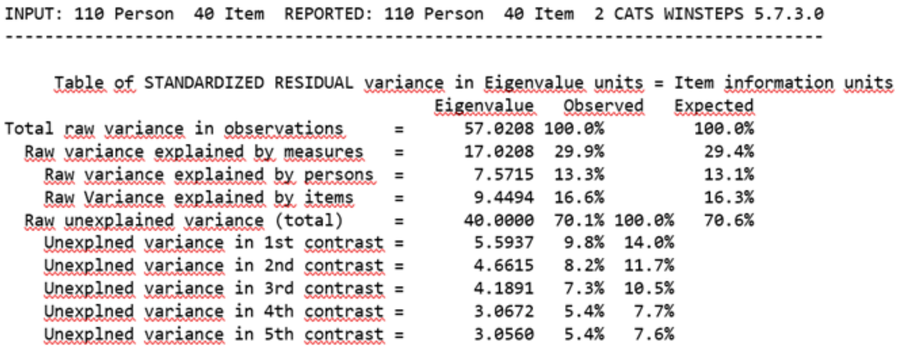
Construct Validity

Construct validity can be measured by referring to polarity items. Item polarity refers to the observed Point Measure Correlation (PT-MEA Corr) value in the range of 0.20 to 0.71. Therefore, it can be concluded that the items are valid. In addition to polarity items, construct validity can be seen through Principal Component Analysis (PCA) to examine items that are in accordance with unidimensionality construct. The results of construct validity test are in Table 6.

Based on data in Table 6, the raw variance explained by the measures is 29.9% (17.0208 eigenvalues), suggesting that the model explains approximately one-third of total variance. Although this percentage aligns closely with the expected variance (29.4%), it indicates that the instrument has limited capacity in capturing the full dimensionality of underlying construct. The variance explained by persons (13.3%) and items (16.6%) indicates a moderate contribution to the overall model structure, but remains below optimal threshold for strong construct representation.

Furthermore, the unexplained variance in the first contrast is 9.8% (5.5937 eigenvalues), which exceeds the commonly accepted threshold of 2.0 eigenvalues for unidimensionality (Bao et al., 2018; Davis & Boone, 2021). This suggests the possible presence of secondary dimensions or noise within the instrument (Köhler et al., 2020). Subsequent contrasts (2nd to 5th) also contribute between 5.4% and 8.1% to the unexplained variance, which reinforces the suspicion of multidimensionality and construct instability.

Table 6. Results of Instrument Construct Validity Test



```
INPUT: 110 Person  40 Item  REPORTED: 110 Person  40 Item  2 CATS WINSTEPS 5.7.3.0
--------------------------------------------------------------------------------

    Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units
                                          Eigenvalue   Observed   Expected
Total raw variance in observations     =      57.0208 100.0%         100.0%
  Raw variance explained by measures   =      17.0208  29.9%          29.4%
    Raw variance explained by persons  =       7.5715  13.3%          13.1%
    Raw Variance explained by items    =       9.4494  16.6%          16.3%
  Raw unexplained variance (total)     =      40.0000  70.1% 100.0%   70.6%
    Unexplned variance in 1st contrast =       5.5937   9.8%  14.0%
    Unexplned variance in 2nd contrast =       4.6615   8.2%  11.7%
    Unexplned variance in 3rd contrast =       4.1891   7.3%  10.5%
    Unexplned variance in 4th contrast =       3.0672   5.4%   7.7%
    Unexplned variance in 5th contrast =       3.0560   5.4%   7.6%
```

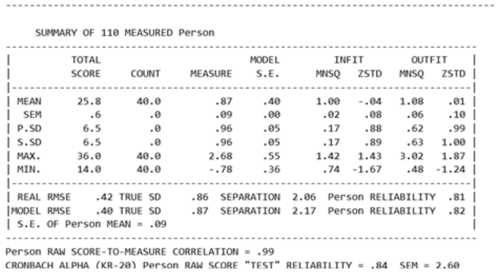
Based on data in Table 6, the raw variance explained by the measures is 29.9% (17.0208 eigenvalues), suggesting that model explains approximately one-third of total variance. Although this percentage aligns closely with the expected variance (29.4%), it indicates that instrument has limited capacity in capturing the full dimensionality of the underlying construct. The variance explained by persons (13.3%) and items (16.6%) indicates a moderate contribution to the overall model structure, but remains below the optimal threshold for strong construct representation.

Furthermore, the unexplained variance in the first contrast is 9.8% (5.5937 eigenvalues), which exceeds the commonly accepted threshold of 2.0 eigenvalues for unidimensionality (Bao et al., 2018; Davis & Boone, 2021). This suggests the possible presence of secondary dimensions or noise within the instrument (Köhler et al., 2020). Subsequent contrasts (2nd to 5th) also contribute between 5.4% and 8.1% to the unexplained variance, which reinforces the suspicion of multidimensionality and construct instability.

These findings indicate that construct validity of instrument may be inadequate. An ideal Rasch-based instrument should explain more than 40% of the raw variance and produce a first contrast eigenvalue below 2.0 to support unidimensionality. In this case, neither criterion is sufficiently met. Therefore, it is advisable to revise instrument, either by refining existing items or by including sub-constructs that align more closely with theoretical indicators (Arfiani et al., 2023; Rafatbakhsh et al., 2021). The presence of unexplained variance in multiple contrasts also suggests possibility of item redundancy or poorly defined dimensions, which should be addressed through iterative testing and expert validation (Susongko et al., 2024; Tennant & Küçükdeveci, 2023).

Reliability

The personal reliability test produced a person reliability of 0.84 in the good category, meaning that prospective science teachers have good internal consistency in answering the test items. The item fit of the person reliability test is presented in Figure 2.



```
INPUT: 110 Person  40 Item  REPORTED: 110 Person  40 Item  2 CATS WINSTEPS 5.7.3.0
--------------------------------------------------------------------------------

     SUMMARY OF 110 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT    MEASURE    S.E.       MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      25.8      40.0        .87     .40       1.00   -.04   1.08    .01 |
|  SEM        .6        .0        .09     .00        .02    .08    .06    .10 |
| P.SD       6.5        .0        .96     .05        .17    .88    .62    .99 |
| S.SD       6.5        .0        .96     .05        .17    .89    .63   1.00 |
| MAX.      36.0      40.0       2.68     .55       1.42   1.43   3.02   1.87 |
| MIN.      14.0      40.0       -.78     .36        .74  -1.67    .48  -1.24 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .42 TRUE SD     .86  SEPARATION  2.06  Person RELIABILITY  .81 |
|MODEL RMSE    .40 TRUE SD     .87  SEPARATION  2.17  Person RELIABILITY  .82 |
| S.E. OF Person MEAN = .09                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .84  SEM = 2.60
```

Figure 2. Person Reliability Test Results

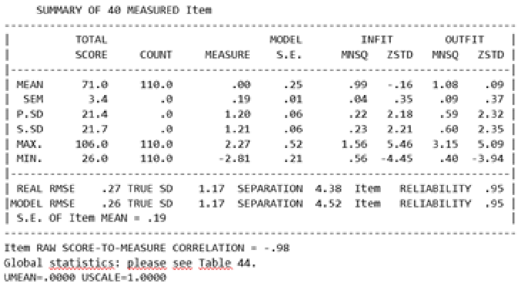
The results of the person reliability analysis on this instrument show a Person Reliability value of 0.92, which is included in the very high category. This value indicates that the instrument has very good consistency in measuring scientific reasoning abilities between individuals (Kusuma et al., 2022; Tennant & Küçükdeveci, 2023). This high reliability indicates that if participants are retested under similar conditions, the measurement results will tend to be stable and reproducible.

In addition, the Person Separation Index (PSI) was recorded at 3.26, which indicates that this instrument is able to differentiate individuals into more than three levels of ability. PSI value above 2.0 is sufficient to separate participants into three levels of ability, and the higher the PSI value, the better the instrument's discrimination power. This shows that the instrument can separate respondents based on their scientific reasoning abilities with good precision.

The Cronbach's Alpha (KR-20) value of 0.99 further strengthens that the internal consistency between the test items is very high. This indicates that the items in the instrument support each other in measuring the same construct, namely, scientific reasoning. This value even exceeds the minimum threshold of high reliability, which is 0.70, and is higher than the minimum recommendation of 0.80 in the context of educational assessment (Freed et al., 2022).

However, even though person reliability is high, it should be noted that the standard error of measurement (SEM) value of ±0.87 must also be taken into account as the limit of estimation accuracy. SEM that is too large can cause doubt in individual interpretation if it is not balanced with the number of items and the quality of the questions (Adu et al., 2023).

The reliability testing of the test items can be seen in Figure 3, which produces an item reliability value of 0.95, meaning it has very good reliability. The Cronbach's Alpha Coefficient value of 0.84 indicates that the test instrument has a very high level of internal consistency.



```
     SUMMARY OF 40 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT    MEASURE    S.E.       MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      71.0     110.0        .00     .25        .99   -.16   1.08    .09 |
|  SEM       3.4        .0        .19     .01        .04    .35    .09    .37 |
| P.SD      21.4        .0       1.20     .06        .22   2.18    .59   2.32 |
| S.SD      21.7        .0       1.21     .06        .23   2.21    .60   2.35 |
| MAX.     106.0     110.0       2.27     .52       1.56   5.46   3.15   5.09 |
| MIN.      26.0     110.0      -2.81     .21        .56  -4.45    .40  -3.94 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .27 TRUE SD    1.17  SEPARATION  4.38  Item   RELIABILITY  .95 |
|MODEL RMSE    .26 TRUE SD    1.17  SEPARATION  4.52  Item   RELIABILITY  .95 |
| S.E. OF Item MEAN = .19                                                     |
-------------------------------------------------------------------------------
Item RAW SCORE-TO-MEASURE CORRELATION = -.98
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000
```

Figure 3. Results of the Reliability Test of Scientific Reasoning Skills Items

The results of the item reliability analysis showed a value of 0.95, which is included in the very high category. This value indicates that the order of the difficulty level of the items in the instrument is very consistent. This means that if this instrument is used on different groups of respondents, it is likely that the arrangement or ranking of the item difficulties will remain stable. High item reliability indicates that all questions contribute strongly and consistently to the measurement of the scientific reasoning construct. An item reliability value of ≥0.90 indicates excellent measurement quality and supports the construct validity of the instrument. High reliability value reflects the stability of the instrument in compiling a hierarchy of items in linear and meaningful units of measurement. In addition to reliability, the analysis also revealed that the item separation value reached 4.32, indicating that this instrument can differentiate items into at least four to five levels of difficulty. This indicates that this measuring instrument is very sensitive in identifying variations in the level of difficulty of each question compiled. A separation value of ≥3.0 indicates that the items in instrument have an adequate distribution of difficulty, and the instrument can measure various levels of individual ability sharply. In other words, the higher the item separation, the higher the instrument's ability to cover a wide spectrum of participant skills (Sovey et al., 2022).

This finding is supported by the study of Sofwan et al. (2022), who used the Rasch model to develop a scientific reasoning instrument and stated that high reliability and separation values are important indicators that the construct being measured is valid, structured, and representative. They emphasize the importance of these two indicators as a basis for stating that an instrument is truly able to differentiate respondents based on abilities to be measured (Muslihin et al., 2022).

Overall, the reliability and separation values in this instrument indicate that the instrument has very good psychometric quality in the context of measuring scientific reasoning. However, to improve the accuracy of interpretation and ensure the generalization of the instrument to various contexts, it is recommended to conduct further validation tests, such as Differential Item Functioning (DIF) analysis and unidimensionality tests. This cross-validation is important to ensure that no items are biased and that all items are actually measuring same construct consistently.

Test Item Characteristics

1. Level of Difficulty

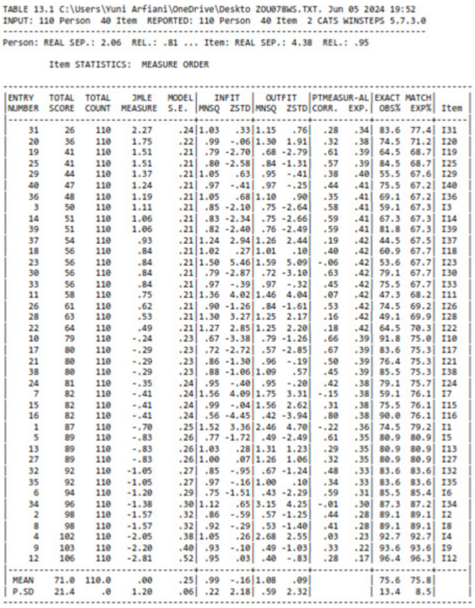
The characteristics of analyzed items are the level of difficulty of items using Item Measure on the WINSTEPS Version 5.7.3.0 application. The standard deviation (SD) value in this test is 1.20, so the category of Level of Difficulty of Questions is based on the Logit Value in Table 6.

Table 6. Category of Question Item Difficulty Level based on Logit Value

Logit Value	Category
More than +1.20 SD	Very Difficult
0.0 logit +1.20 SD	Difficult
0.0 logit -1.20 SD	Medium
Less than -1.20 SD	Easy

The logit value located in the JMLE MEASURE column indicates the level of difficulty of each item (Figure 4). Positive values indicate lower difficulty (easier), while negative values indicate higher difficulty (more difficult). The level of difficulty of the items is sorted from top to bottom, from the most difficult to the easiest. Analysis of item difficulty levels based on MEASURE logit shows that the range of logit values is between -1.25 to +1.28, which indicates that the variation in question difficulty in the instrument is quite good. Negative logit values indicate items that are more difficult because they can only be answered by participants with high abilities, while positive values indicate items that are relatively easier and can be answered by participants with lower abilities. The order of items arranged from the highest to the lowest logit reflects the hierarchy of difficulty levels, from the easiest to the most challenging items. In the context of the Rasch model, item logits are ideally evenly distributed in order to measure various levels of student ability representatively, and not accumulate at just one level. The proportional distribution of logit values, as shown in Figure 4, supports the validity of the instrument's content because it shows that the instrument covers a range of scientific reasoning abilities from simple to complex. An even distribution of difficulty levels will increase the sensitivity of the instrument in identifying individual ability levels, as well as preventing ceiling effects (items that are too easy) or floor effects (items that are too difficult).

However, some items in the table show extreme values, either too easy or too difficult. For example, the item with the highest logit of +1.28 is classified as very easy, while the item with a logit of -1.25 is the most difficult. Items with extreme logits are recommended for re-evaluation because they may not make the maximum contribution to mapping participant abilities. Items that are too easy tend to be uninformative in distinguishing participants with high abilities, while items that are too difficult can reduce motivation and response validity.



```
TABLE 13.1 C:\Users\Yuni Arfiani\OneDrive\Deskto ZOU078WS.TXT. Jun 05 2024 19:52
INPUT: 110 Person  40 Item  REPORTED: 110 Person  40 Item  2 CATS WINSTEPS 5.7.3.0
--------------------------------------------------------------------------------
Person: REAL SEP.: 2.06  REL.: .81 ... Item: REAL SEP.: 4.38  REL.: .95

        Item STATISTICS:  MEASURE ORDER

------------------------------------------------------------------------------------
|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 |
|------------------------------------+----------+----------+-----------+-----------+------|
|    31     26    110     2.27    .24|1.03   .33|1.15   .76|  .28   .34| 83.6  77.4| I31  |
|    20     36    110     1.75    .22| .99  -.06|1.30  1.91|  .32   .38| 74.5  71.2| I20  |
|    19     41    110     1.51    .21| .79 -2.70| .68 -2.79|  .61   .39| 64.5  68.7| I19  |
|    25     41    110     1.51    .21| .80 -2.58| .84 -1.31|  .57   .39| 84.5  68.7| I25  |
|    29     44    110     1.37    .21|1.05   .63| .95  -.41|  .38   .40| 55.5  67.6| I29  |
|    40     47    110     1.24    .21| .97  -.41| .97  -.25|  .44   .41| 75.5  67.2| I40  |
|    36     48    110     1.19    .21|1.05   .68|1.10   .90|  .35   .41| 69.1  67.2| I36  |
|     3     50    110     1.11    .21| .85 -2.10| .75 -2.64|  .58   .41| 59.1  67.3| I3   |
|    14     51    110     1.06    .21| .83 -2.34| .75 -2.66|  .59   .41| 67.3  67.3| I14  |
|    39     51    110     1.06    .21| .82 -2.40| .76 -2.49|  .59   .41| 81.8  67.3| I39  |
|    37     54    110      .93    .21|1.24  2.94|1.26  2.44|  .19   .42| 44.5  67.5| I37  |
|    18     56    110      .84    .21|1.02   .27|1.01   .10|  .40   .42| 60.9  67.7| I18  |
|    23     56    110      .84    .21|1.50  5.46|1.59  5.09| -.06   .42| 53.6  67.7| I23  |
|    30     56    110      .84    .21| .79 -2.87| .72 -3.10|  .63   .42| 79.1  67.7| I30  |
|    33     56    110      .84    .21| .97  -.39| .97  -.32|  .45   .42| 75.5  67.7| I33  |
|    11     58    110      .75    .21|1.36  4.02|1.46  4.04|  .07   .42| 47.3  68.2| I11  |
|    26     61    110      .62    .21| .90 -1.26| .84 -1.61|  .53   .42| 74.5  69.2| I26  |
|    28     63    110      .53    .21|1.30  3.27|1.25  2.17|  .16   .42| 49.1  69.9| I28  |
|    22     64    110      .49    .21|1.27  2.85|1.25  2.20|  .18   .42| 64.5  70.3| I22  |
|    10     79    110     -.24    .23| .67 -3.38| .79 -1.26|  .66   .39| 91.8  75.0| I10  |
|    17     80    110     -.29    .23| .72 -2.72| .57 -2.85|  .67   .39| 83.6  75.3| I17  |
|    21     80    110     -.29    .23| .86 -1.30| .96  -.19|  .50   .39| 76.4  75.3| I21  |
|    38     80    110     -.29    .23| .88 -1.06|1.09   .57|  .45   .39| 85.5  75.3| I38  |
|    24     81    110     -.35    .24| .95  -.40| .95  -.20|  .42   .38| 79.1  75.7| I24  |
|     7     82    110     -.41    .24|1.56  4.09|1.75  3.31| -.15   .38| 59.1  76.1| I7   |
|    15     82    110     -.41    .24| .99  -.04|1.56  2.62|  .31   .38| 75.5  76.1| I15  |
|    16     82    110     -.41    .24| .56 -4.45| .42 -3.94|  .80   .38| 90.0  76.1| I16  |
|     1     87    110     -.70    .25|1.52  3.36|2.46  4.70| -.22   .36| 74.5  79.2| I1   |
|     5     89    110     -.83    .26| .77 -1.72| .49 -2.49|  .61   .35| 80.9  80.9| I5   |
|    13     89    110     -.83    .26|1.03   .28|1.31  1.23|  .29   .35| 80.9  80.9| I13  |
|    27     89    110     -.83    .26|1.00   .07|1.26  1.06|  .32   .35| 80.9  80.9| I27  |
|    32     92    110    -1.05    .27| .85  -.95| .67 -1.24|  .48   .33| 83.6  83.6| I32  |
|    35     92    110    -1.05    .27| .97  -.16|1.00   .10|  .34   .33| 83.6  83.6| I35  |
|     6     94    110    -1.20    .29| .75 -1.51| .43 -2.29|  .59   .31| 85.5  85.4| I6   |
|    34     96    110    -1.38    .30|1.12   .65|3.15  4.25| -.01   .30| 87.3  87.2| I34  |
|     2     98    110    -1.57    .32| .86  -.59| .57 -1.25|  .44   .28| 89.1  89.1| I2   |
|     8     98    110    -1.57    .32| .92  -.29| .53 -1.40|  .41   .28| 89.1  89.1| I8   |
|     4    102    110    -2.05    .38|1.05   .26|2.68  2.55|  .03   .23| 92.7  92.7| I4   |
|     9    103    110    -2.20    .40| .93  -.10| .49 -1.03|  .33   .22| 93.6  93.6| I9   |
|    12    106    110    -2.81    .52| .95   .03| .40  -.83|  .28   .17| 96.4  96.3| I12  |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    71.0  110.0      .00    .25| .99  -.16|1.08   .09|           | 75.6  75.8|      |
| P.SD    21.4     .0     1.20    .06| .22  2.18| .59  2.32|           | 13.4   8.5|      |
------------------------------------------------------------------------------------
```

Figure 4. Results of the Scientific Reasoning Instrument Item Measure

To maintain the effectiveness of the measurement, it is important to maintain an even distribution of difficulty levels that are relevant to the indicators of the construct to be measured. Revisions to items with extreme logits can be made by considering cognitive structure required by participants to answer them, as well as re-evaluating the relevance of the item content to the indicators of scientific reasoning. Thus, this instrument can continue to be developed to provide valid, sensitive, and meaningful measurement results.

Table 7 allows us to identify items that are too difficult or too easy. Items with unsatisfactory fit should be reviewed because they may not measure what they are supposed to measure or may confuse respondents. This is crucial to ensure that the measurement of the desired ability in psychometrics or education is both valid and reliable. The results of the analysis of the level of difficulty of the items based on the logit value and its category in Table 6 are presented in Table 7.

Table 7. Level of Difficulty of Scientific Reasoning Test Instrument Items

Item Number	Total Score	Total Count	JMLE Measure	Level of Difficulty
31	26	110	2.27	Very Difficult
20	36	110	1.75	Very Difficult
19	41	110	1.51	Very Difficult
25	41	110	1.51	Very Difficult
29	44	110	1.37	Very Difficult
40	47	110	1.24	Very Difficult
36	48	110	1.19	Difficult
3	50	110	1.11	Difficult
14	51	110	1.06	Difficult
39	51	110	1.06	Difficult
37	54	110	0.93	Difficult
18	56	110	0.84	Difficult
23	56	110	0.84	Difficult
30	56	110	0.84	Difficult
33	56	110	0.84	Difficult
11	58	110	0.75	Difficult
26	61	110	0.62	Difficult
28	63	110	0.53	Difficult
22	64	110	0.49	Difficult
10	79	110	-0.24	Medium
17	80	110	-0.29	Medium
21	80	110	-0.29	Medium
38	80	110	-0.29	Medium
24	81	110	-0.35	Medium
7	82	110	-0.41	Medium

Item Number	Total Score	Total Count	JMLE Measure	Level of Difficulty
15	82	110	-0.41	Medium
16	82	110	-0.41	Medium
1	87	110	-0.70	Medium
5	89	110	-0.83	Medium
13	89	110	-0.83	Medium
27	89	110	-0.83	Medium
32	92	110	-1.05	Medium
35	92	110	-1.05	Medium
6	94	110	-1.20	Medium
34	96	110	-1.38	Medium
2	98	110	-1.57	Medium
8	98	110	-1.57	Medium
4	102	110	-2.05	Medium
9	103	110	-2.20	Medium
12	106	110	-2.81	Medium

Based on Table 7, the number of questions can be grouped according to the level of difficulty as presented in Figure 5.

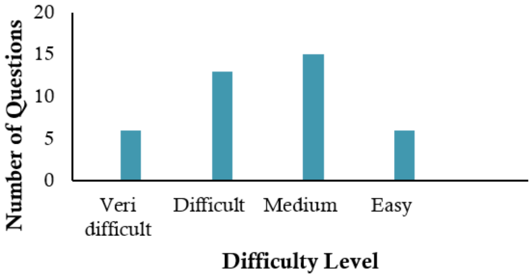


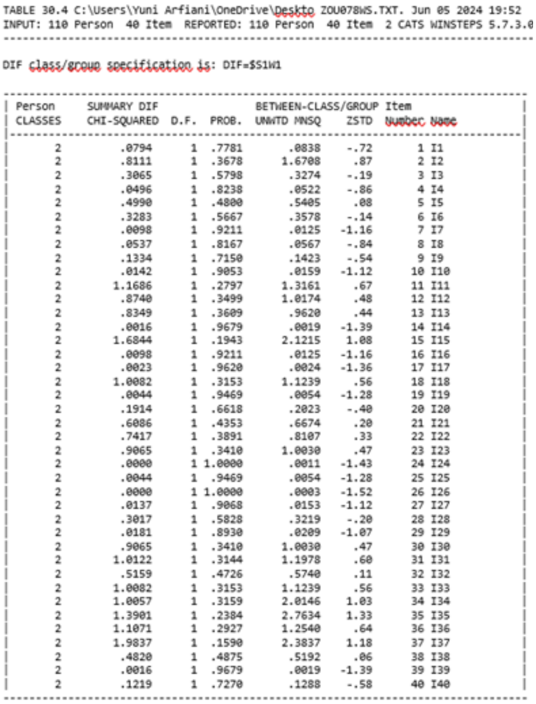
Figure 5. Number of Questions Based on Difficulty Level

Figure 5 shows the distribution of the number of questions based on the level of difficulty, which is categorized into four levels, namely Very Difficult, Difficult, Medium, and Easy. Based on the graph, the majority of questions are included in the Medium category, with 16 items, followed by the Difficult category with 13 items. Meanwhile, the Very Difficult and Easy categories each consist of only 6 questions. This distribution indicates that the compilation of questions tends to focus on moderate to difficult levels of difficulty, which is the right strategy in the context of measuring scientific reasoning ability. A good instrument should have an even distribution of difficulty levels, with a dominance of medium and difficult items to optimally reach the variation in participant abilities. Medium items serve as measurement anchors, while difficult and very difficult items help detect high abilities, and easy items detect the lower limit of ability.

However, the proportion of items in the Very Difficult and Easy categories that are both low needs to be considered. This imbalance can reduce the sensitivity of the instrument in measuring the abilities of participants who are at the ends of the spectrum—either very high or very low. The dominance of questions at only one level of difficulty has the potential to cause a ceiling effect or floor effect, namely a condition where very high or very low participants cannot be distinguished accurately by the instrument (Karim et al., 2021). Therefore, although the arrangement of items in this instrument is quite proportional, further development is recommended to add item variations, especially in categories that have not been optimally represented. This is important so that the instrument can accommodate participants with a wider range of abilities and provide comprehensive, valid, and reliable measurement results.

2. Item DIF

One of the indicators of valid measurement is that the instruments and items used do not contain bias. An instrument is said to be biased if there is an individual with certain characteristics who is more advantaged than an individual with other characteristics. Items are detected as DIF if they have a probability value of less than 5%. In analyzing DIF using Figure 6.



```
TABLE 30.4 C:\Users\Yuni Arfiani\OneDrive\Deskto ZOU078WS.TXT. Jun 05 2024 19:52
INPUT: 110 Person  40 Item  REPORTED: 110 Person  40 Item  2 CATS WINSTEPS 5.7.3.0
--------------------------------------------------------------------------------

DIF class/group specification is: DIF=$S1W1

-------------------------------------------------------------------------------
| Person    SUMMARY DIF                BETWEEN-CLASS/GROUP Item                 |
| CLASSES   CHI-SQUARED  D.F.  PROB.  UNWTD MNSQ    ZSTD  Number Name           |
|-----------------------------------------------------------------------------|
|       2        .0794      1  .7781       .0838    -.72       1 I1             |
|       2        .8111      1  .3678      1.6708     .87       2 I2             |
|       2        .3065      1  .5798       .3274    -.19       3 I3             |
|       2        .0496      1  .8238       .0522    -.86       4 I4             |
|       2        .4990      1  .4800       .5405     .08       5 I5             |
|       2        .3283      1  .5667       .3578    -.14       6 I6             |
|       2        .0098      1  .9211       .0125   -1.16       7 I7             |
|       2        .0537      1  .8167       .0567    -.84       8 I8             |
|       2        .1334      1  .7150       .1423    -.54       9 I9             |
|       2        .0142      1  .9053       .0159   -1.12      10 I10            |
|       2       1.1686      1  .2797      1.3161     .67      11 I11            |
|       2        .8740      1  .3499      1.0174     .48      12 I12            |
|       2        .8349      1  .3609       .9620     .44      13 I13            |
|       2        .0016      1  .9679       .0019   -1.39      14 I14            |
|       2       1.6844      1  .1943      2.1215    1.08      15 I15            |
|       2        .0098      1  .9211       .0125   -1.16      16 I16            |
|       2        .0023      1  .9620       .0024   -1.36      17 I17            |
|       2       1.0082      1  .3153      1.1239     .56      18 I18            |
|       2        .0044      1  .9469       .0054   -1.28      19 I19            |
|       2        .1914      1  .6618       .2023    -.40      20 I20            |
|       2        .6086      1  .4353       .6674     .20      21 I21            |
|       2        .7417      1  .3891       .8107     .33      22 I22            |
|       2        .9065      1  .3410      1.0030     .47      23 I23            |
|       2        .0000      1 1.0000       .0011   -1.43      24 I24            |
|       2        .0044      1  .9469       .0054   -1.28      25 I25            |
|       2        .0000      1 1.0000       .0003   -1.52      26 I26            |
|       2        .0137      1  .9068       .0153   -1.12      27 I27            |
|       2        .3017      1  .5828       .3219    -.20      28 I28            |
|       2        .0181      1  .8930       .0209   -1.07      29 I29            |
|       2        .9065      1  .3410      1.0030     .47      30 I30            |
|       2       1.0122      1  .3144      1.1978     .60      31 I31            |
|       2        .5159      1  .4726       .5740     .11      32 I32            |
|       2       1.0082      1  .3153      1.1239     .56      33 I33            |
|       2       1.0057      1  .3159      2.0146    1.03      34 I34            |
|       2       1.3901      1  .2384      2.7634    1.33      35 I35            |
|       2       1.1071      1  .2927      1.2540     .64      36 I36            |
|       2       1.9837      1  .1590      2.3837    1.18      37 I37            |
|       2        .4820      1  .4875       .5192     .06      38 I38            |
|       2        .0016      1  .9679       .0019   -1.39      39 I39            |
|       2        .1219      1  .7270       .1288    -.58      40 I40            |
-------------------------------------------------------------------------------
```

Figure 6. DIF Item Test Results

If we look at the probability column in Figure 6, there are no items with a probability value below 0.05 (5%). This means that there is no statistically significant difference in how respondents from two different groups (e.g., class or certain characteristic groups) answer the questions. Thus, it can be concluded that there are no items in the scientific reasoning skills instrument that show bias towards a particular group, or in Rasch terms, known as the absence of uniform DIF.

DIF analysis is an important part of instrument validation because it allows testing the fairness of the measurement. Items are said to contain bias if the difference in participants' answers is not caused by ability, but by group characteristics such as gender, educational background, or class (Chanpleng et al., 2015; Lu et al., 2021; Zumbo et al., 2016). Items that have a probability value below 0.05 in the Chi-Square analysis and a |DIF contrast| value of more than ±0.5 logit are suspected of containing potential bias. In this data, no items meet these criteria, thus supporting that the instrument works fairly and evenly across all respondent groups.

The absence of DIF is an indicator that all participants, regardless of group differences, have an equal opportunity to answer each item correctly according to their ability level (Chanpleng et al., 2015; Huang et al., 2016). This is especially important for scientific reasoning instruments, because this construct is closely related to high-level thinking skills that must be equally accessible to all participants (Krell et al., 2020; Susongko et al., 2021). In addition, this finding also confirms that the process of developing and compiling items has been carried out by paying attention to aspects of fairness and conceptual representation. With no indication of bias between groups, it can be concluded that this instrument is suitable for use in diverse educational contexts, both in terms of student background and other class characteristics.

CONCLUSION

Based on the results of the validity and reliability analysis, there were 25 questions that were declared valid. The valid items covered all indicators of scientific reasoning ability, namely variable control, combinatorial thinking, probability thinking, relational thinking, and proportional thinking. The test instrument also had a very high level of internal consistency. The results of person reliability showed that prospective science teachers had good internal consistency in answering questions (0.84), while the results of item reliability showed that the questions had very good reliability (0.95). The results of the level of difficulty of the questions showed a balanced distribution between very difficult, difficult, moderate, and easy questions. The analysis of the level of difficulty showed that several questions needed to be reviewed to ensure that the questions were neither too easy nor too difficult. The DIF analysis showed that no questions were detected to have bias. This means that the instrument developed is fair and can be used to measure scientific reasoning ability without favoring one group of individuals based on certain characteristics. Overall, the scientific reasoning ability test instrument developed is proven to be valid and reliable for use on prospective science teachers. The use of this instrument can help educators in evaluating students' scientific reasoning abilities, providing constructive feedback, and improving the quality of learning in science. Further analysis and revision are needed to improve the effectiveness of this instrument.

ACKNOWLEDGMENT

The author would like to thank the Indonesian Education Scholarship, the Center for Financing and Assessment of Higher Education of the Ministry of Education and Technology, and the Education Fund Management Institute as the funding providers for this research.

REFERENCES

Abate, T., Michael, K., & Angell, C. (2020). Assessment of Scientific Reasoning: Development and Validation of Scientific Reasoning Assessment Tool. *Eurasia Journal of Mathematics, Science and Technology Education*, 16(12). https://doi.org/10.29333/ejmste/9353

Adu, P., Popoola, T., Roemer, A., Collings, S., Aspin, C., Medvedev, O. N., & Simpson, C. R. (2023). Validation and Cultural Adaptation of the Motors of COVID-19 Vaccination Acceptance Scale (MoVac-COVID19S) in German. *Psychological Test Adaptation and Development*, 4(1). https://doi.org/10.1027/2698-1866/a000064

Anjani, F., Supeno, S., & Subiki, S. (2020). Kemampuan Penalaran Ilmiah Siswa SMA dalam Pembelajaran Fisika Menggunakan Model Inkuiri Terbimbing disertai Diagram Berpikir Multidimensi. *Lantanida Journal*, 8(1). https://doi.org/10.22373/lj.v8i1.6306

Arfiani, Y., Susongko, P., & Kusuma, M. (2023). Construct validity analysis with messick validity approach and rasch model application on scientific reasoning test items. *THABIEA : JOURNAL OF NATURAL SCIENCE TEACHING*, 6(1). https://doi.org/10.21043/thabiea.v6i1.18918

Bao, L., Koenig, K., Xiao, Y., Fritchman, J., Zhou, S., & Chen, C. (2022). Theoretical model and quantitative assessment of scientific thinking and reasoning. *Physical Review Physics Education Research*, 18(1). https://doi.org/10.1103/PhysRevPhysEducRes.18.010115

Bao, L., Xiao, Y., Koenig, K., & Han, J. (2018). Validity evaluation of the Lawson classroom test of scientific reasoning. *Physical Review Physics Education Research*, 14(2). https://doi.org/10.1103/PhysRevPhysEducRes.14.020106

Ben-Horin, H., Kali, Y., & Tal, T. (2023). The Fifth Dimension in Socio-Scientific Reasoning: Promoting Decision-Making about Socio-Scientific Issues in a Community. *Sustainability (Switzerland),* 15(12). https://doi.org/10.3390/su15129708

Chanpleng, P., Lawthong, N., & Ngudgratoke, S. (2015). A study of Quality Assessment of Science Instructional Management in Thailand: An analysis of Differential Item Functioning and Test Functioning in Mixed format Tests. *Procedia - Social and Behavioral Sciences*, 191. https://doi.org/10.1016/j.sbspro.2015.04.348

Davis, D. R., & Boone, W. (2021). Using Rasch analysis to evaluate the psychometric functioning of the other-directed, lighthearted, intellectual, and whimsical (OLIW) adult playfulness scale. *International Journal of Educational Research Open*, 2. https://doi.org/10.1016/j.ijedro.2021.100054

Díaz, C., Dorner, B., Hussmann, H., & Strijbos, J. W. (2023). Conceptual review on scientific reasoning and scientific thinking. *Current Psychology,* 42(6). https://doi.org/10.1007/s12144-021-01786-5

Firdaus, F., Wiyanto, W., Putra, N. M. D., & Isnaen, W. (2025). Design of instruments for scientific creative thinking skills and creative thinking digital skills: Rasch models and confirmatory factor analysis. *Eurasia Journal of Mathematics, Science and Technology Education*, 21(5). https://doi.org/10.29333/ejmste/16310

Freed, R., McKinnon, D., Fitzgerald, M., & Norris, C. M. (2022). Development and validation of an astronomy self-efficacy instrument for understanding and doing. *Physical Review Physics Education Research*, 18(1). https://doi.org/10.1103/PhysRevPhysEducRes.18.010117

Fütterer, T., Steinhauser, R., Zitzmann, S., Scheiter, K., Lachner, A., & Stürmer, K. (2023). Development and validation of a test to assess teachers' knowledge of how to operate technology. *Computers and Education Open*, 5, 100152. https://doi.org/10.1016/j.caeo.2023.100152

Hadarah, H., & Tulhikmah, R. (2019). Epistemology Philosophy Development of Assessment Instruments Student Chemistry Learning Outcomes. *Journal for the Education of Gifted Young Scientists*, 7(3). https://doi.org/10.17478/jegys.605097

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology,* 36(2). https://doi.org/10.1080/01443410.2014.946890

Jufri, A. W., Setiadi, D., & Sripatmi. (2016). Scientific reasoning ability of prospective student teacher in the excellence program of mathematics and science teacher education in University of Mataram. *Jurnal Pendidikan IPA Indonesia,* 5(1). https://doi.org/10.15294/jpii.v5i1.5792

Jumadi, J., Sukarelawan, M. I., & Kuswanto, H. (2023). An investigation of item bias in the four-tier diagnostic test using Rasch model. *International Journal of Evaluation and Research in Education*, 12(2), 622–629. https://doi.org/10.11591/ijere.v12i2.22845

Kara, S., & Aslan, O. (2024). Evaluation of the effect of in-class activity-based practices on scientific reasoning skills of pre-service science teachers. *International Journal of Science Education,* 46(8). https://doi.org/10.1080/09500693.2023.2258253

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature and Culture*, 6(2). https://doi.org/10.30659/e.6.2.256-269

Krell, M., Mathesius, S., van Driel, J., Vergara, C., & Krüger, D. (2020). Assessing scientific reasoning competencies of pre-service science teachers: translating a German multiple-choice instrument into English and Spanish. *International Journal of Science Education*, 42(17). https://doi.

org/10.1080/09500693.2020.1837989

Kusuma, I. Y., Triwibowo, D. N., Pratiwi, A. D. E., & Pitaloka, D. A. E. (2022). Rasch Modelling to Assess Psychometric Validation of the Knowledge about Tuberculosis Questionnaire (KATUB-Q) for the General Population in Indonesia. *International Journal of Environmental Research and Public Health*, 19(24). https://doi.org/10.3390/ijerph192416753

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1). https://doi.org/10.1002/tea.3660150103

Lu, R., Guo, H., & Dorans, N. J. (2021). Robustness of Weighted Differential Item Functioning (DIF) Analysis: The Case of Mantel–Haenszel DIF Statistics. *ETS Research Report Series*, 2021(1). https://doi.org/10.1002/ets2.12325

Maulana, S., Rusilowati, A., Nugroho, S. E., & Susilaningsih, E. (2023). Implementasi Rasch Model dalam Pengembangan Instrumen Tes Diagnostik. *Prosiding Seminar Nasional Pascasarjana (PROSNAMPAS)*, 6(1).

Muslihin, H. Y., Suryana, D., Ahman, Suherman, U., & Dahlan, T. H. (2022). Analysis of the Reliability and Validity of the Self-Determination Questionnaire Using Rasch Model. *International Journal of Instruction,* 15(2), 207–222. https://doi.org/10.29333/iji.2022.15212a

Nugroho, S. E., & Waslam. (2020). Physics experiment activities to stimulate interest in learning physics and reasoning in high school students. *Journal of Physics: Conference Series,* 1567(2). https://doi.org/10.1088/1742-6596/1567/2/022069

Rafatbakhsh, E., Ahmadi, A., Moloodi, A., & Mehrpour, S. (2021). Development and Validation of an Automatic Item Generation System for English Idioms. *Educational Measurement: Issues and Practice*, 40(2). https://doi.org/10.1111/emip.12401

Sovey, S., Osman, K., & Matore, M. E. E. M. (2022). Rasch Analysis for Disposition Levels of Computational Thinking Instrument Among Secondary School Students. *Eurasia Journal of Mathematics, Science and Technology Education*, 18(3). https://doi.org/10.29333/ejmste/11794

Suaidah, H. L., Susantini, E., & Hariyono, E. (2023). Determining Learning Activities to Promote Scientific Reasoning in Science Learning: A Literature Review. *IJORER : International Journal of Recent Educational Research*, 4(3). https://doi.org/10.46245/ijorer.v4i3.285

Susongko, P., Arfiani, Y., & Kusuma, M. (2021). Determination of Gender Differential Item Functioning in Tegal-Students' Scientific Literacy Skills with Integrated Science (SLISIS) Test using Rasch Model. *Jurnal Pendidikan IPA Indonesia*, 10(2), 270–281. https://doi.org/10.15294/jpii.v10i2.26775

Susongko, P., Wahab, N. B. A., Arfiani, Y., & Kusuma, M. (2024). Validation and Implementation of 3-Dimensional Scientific Literacy Test (LISA3D Test): Measuring Scientific Literacy for Senior High School Students Based on Scientific Reasoning, Scientific Inquiry, and Nature of Science. *JPII*, 13(3), 459–470. https://doi.org/10.15294/jpii.v13i3.6309

Tennant, A., & Küçükdeveci, A. A. (2023). Application of the Rasch measurement model in rehabilitation research and practice: early developments, current practice, and future challenges. *Frontiers in Rehabilitation Sciences*, 4. https://doi.org/10.3389/fresc.2023.1208670

Thayaseelan, K., Zhai, Y., Li, S., & Liu, X. (2024). Revalidating a measurement instrument of spatial thinking ability for junior and high school students. *Disciplinary and Interdisciplinary Science Education Research*, 6(1). https://doi.org/10.1186/s43031-024-00095-8

Winari, D. R., & Masturi, M. (2023). Pengembangan Instrumen Tes Two-Tier Multiple Choiche Berbasis Hots Pada Materi Rangkaian Arus Bolak-Balik. *INPAFI (Inovasi Pembelajaran Fisika),* 11(01). https://doi.org/10.24114/inpafi.v11i01.43856

Yanto, B. E., Subali, B., & Suyanto, S. (2019). Measurement instrument of scientific reasoning test for biology education students. *International Journal of Instruction*, 12(1). https://doi.org/10.29333/iji.2019.12188a

Yulianti, D., Wiyanto, Rusilowati, A., & Nugroho, S. E. (2020). Student worksheets based on Science, Technology, Engineering and Mathematics (STEM) to facilitate the development of critical and creative thinking skills. *Journal of Physics: Conference Series*, 1567(2). https://doi.org/10.1088/1742-6596/1567/2/022068

Yulianti, E., & Zhafirah, N. N. (2020). Analisis Komprehensif pada Implementasi Pembelajaran dengan Model Inkuiri Terbimbing: Aspek Penalaran Ilmiah. *Jurnal Penelitian Pendidikan IPA*, 6(1). https://doi.org/10.29303/jppipa.v6i1.341

Zhou, S. N., Liu, Q. Y., Koenig, K., Li, Q. Y., Xiao, Y., & Bao, L. (2021). Analysis of two-tier question scoring methods: A case study on the Lawson's classroom test of scientific reasoning. *Journal of Baltic Science Education*, 20(1). https://doi.org/10.33225/jbse/21.20.146

Zulkipli, Z. A., Mohd Yusof, M. M., Ibrahim, N., & Dalim, S. F. (2020). Identifying Scientific Reasoning Skills of Science Education Students. *Asian Journal of University Education*, 16(3). https://doi.org/10.24191/ajue.v16i3.10311

Zumbo, B. D., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. D. (2016). Investigating causal DIF via propensity score methods. *Practical Assessment, Research & Evaluation*, 21(13).