



Development of a Scientific Literacy Assessment Instrument Based on the PISA 2018 Assessment and Analytical Framework at the Elementary Education Level

Hana Yunansah*, Muslim, Ridwan Efendi, Raden Giovanni Ariantara

Universitas Pendidikan Indonesia, Indonesia
Correspondence: E-mail: hanayunansah@upi.edu

ABSTRACT

The low average achievement of Indonesian students' scientific literacy on PISA (Program for International Student Assessment) is thought to be because Indonesian students are not used to solving questions that test reasoning. The facts in the field show that the teacher has not specifically trained students' scientific literacy skills in science learning and the teacher's lack of ability in making assessment instruments to measure students' scientific literacy skills. This indicates the need to develop a scientific literacy assessment instrument that refers to the PISA framework. This study aims to develop a scientific literacy assessment instrument based on the PISA framework on science content in junior high schools and determine the achievement of students' scientific literacy skills. The research method used was R and D research. The instrument in this study was a scientific literacy test item developed based on the 2018 PISA framework which combines aspects of context, knowledge and competence. The simulation-based test and the three-dimensional learning framework generally satisfy logical validity, subject to certain considerations for future test development. The test also demonstrates empirical validity and reliability based on the results of extensive student testing. Therefore, it can be concluded that this developed product is capable of measuring students' cognitive abilities regarding the topic of light waves in accordance with the three-dimensional learning framework.

ARTICLE INFO

Article History:

Submitted/Received 05 Jul 2025

First Revised 02 Aug 2025

Accepted 03 Sep 2025

First Available online 21 Sep 2025

Publication Date 01 Dec 2025

Keyword:

Assessment instrument,
PISA 2018 framework,
Scientific literacy.

1. INTRODUCTION

A paradigm shift in the education system directed toward achieving 21st-century skills in today's challenge-filled globalized era is essential to equip students, particularly at the primary and secondary education levels, to become high-quality human resources. Through these efforts, students are expected to develop their capacity to handle new, more open situations and solve real-world problems. The revised 2013 Curriculum emphasizes that science learning in schools is expected to lead students toward acquiring 21st-century skills. Science learning should ideally be conducted through scientific inquiry to foster the ability to think, work, and adopt scientific attitudes, as well as communicate them as essential life skill aspects. Scientific attitude refers to a set of attitudes that guide individuals to think critically, objectively, and systematically when dealing with scientific problems and processes (Syah et al., 2023).

Science learning at the Junior High School (SMP/MTs) level is viewed not only as a means to transfer science (transfer of science) but also to build higher-order thinking skills (analytical, synthesis, critical, creative, and innovative) through scientific work experiences. Knowledge, skills, thinking abilities, and attitudes derived from science learning will equip students to live in society and pursue further studies, given the characteristics of science as the foundation of various basic and applied sciences. Furthermore, Biology encompasses scientific and religious values, as biological phenomena can be studied as part of the signs of Allah's creation (ayat kauniyah) and integrated with relevant Qur'anic verses and Hadith (Ramadhan et al., 2021).

Regarding science learning assessment, the 2013 Curriculum requires the use of authentic assessment. Paradigmatically, authentic assessment requires the realization of authentic instruction and authentic learning. It is believed that authentic assessment is better able to provide valid and holistic information regarding student capabilities. Assessments are designed to measure knowledge and concepts, process skills, and higher-order reasoning (critical, logical, creative thinking) while adopting item formats similar to PISA and TIMSS. This encourages learning to contribute to improving students' scientific literacy while uncovering scientific, critical, creative, and innovative thinking abilities. The scientific literacy assessment framework developed based on PISA consists of three interrelated aspects: scientific context, scientific knowledge, and scientific competencies (Zhang et al., 2023).

One of the international studies regarding students' scientific literacy skills is PISA (Program for International Student Assessment). This international study was developed by several developed countries under the Organisation for Economic Co-operation and Development (OECD) based in Paris, France, conducted every three years since 2000. Non-OECD countries also partner with the OECD to participate in the PISA study. PISA assesses the knowledge and skills of 15-year-old students across three core domains: reading, mathematics, and science (Pokropek et al., 2022a). PISA aims to evaluate education systems by assessing the extent to which students at the end of their compulsory education can apply their knowledge to real-life situations useful for society.

The 2018 PISA survey results placed Indonesia 74th out of 79 countries, or sixth from the bottom. Based on the PISA results, the average scientific literacy score of Indonesian students decreased from 403 points in 2015 to 396 points in 2018 (Zuhrotul Wardi et al., 2024). The 2018 PISA results indicate that Indonesian students continued to demonstrate low levels of scientific literacy, highlighting the need to strengthen science literacy-oriented learning in Indonesia (Khery et al., 2022), explained that the 2018 PISA results serve as an early warning to transform the educational paradigm in Indonesia. Based on Indonesia's PISA test results,

scientific literacy skills remain unevenly distributed. As many as 40 percent of Indonesian students still perform below the expected minimum capability.

PISA provides data to participating countries to help identify which scientific literacy aspects (competence and knowledge aspects) are well-mastered and which require improvement, serving as a foundation to design learning policies. In addition to scientific literacy surveys, PISA assesses learning, school, and parental aspects through questionnaires. PISA results provide international benchmarks for evaluating student performance and can inform educational policy and decision-making (AlKaabi et al., 2024).

The low scientific literacy performance of Indonesian students in PISA is due to students not being accustomed to solving Higher Order Thinking Skill (HOTS)-based questions. So far, examination questions in Indonesia have had difficulty levels below PISA, which is already HOTS-based. Indonesian students' reasoning abilities remain an important concern, as research has identified variations in their inductive reasoning performance across different levels of difficulty (Soeharto & Csapó, 2022). Therefore, it is necessary to develop assessment instruments to measure students' scientific literacy while training scientific literacy during school science learning to support the improvement of Indonesian students' scientific literacy achievement.

Scientific literacy involves the ability to apply scientific knowledge, including content, procedural, and epistemic knowledge, to explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically (Alotaibi & Alghamdi, 2025). Scientific literacy is multidimensional, involving not only an understanding of science content knowledge but also an understanding of the scientific inquiry process and the rationale behind the processes conducted in science to acquire knowledge.

Scientific literacy encompasses three aspects: context, knowledge, and competencies. Context relates to personal, local/national, and global issues—past or present—that demand an understanding of science and technology. Knowledge relates to understanding facts, concepts, and theories underlying scientific knowledge. Knowledge includes understanding the natural world and technological artifacts (content knowledge), how such ideas are produced (procedural knowledge), and rational understanding of those procedures and justifications for their use (epistemic knowledge). Competencies relate to the ability to offer scientific explanations for phenomena, evaluate and design scientific inquiries, and interpret data and evidence scientifically (Alotaibi & Alghamdi, 2025).

Research on scientific literacy has become a research trend both internationally and nationally. While much research has been conducted on science learning to enhance students' scientific literacy, research directed toward the development of scientific literacy assessment remains limited. Recent research shows that scientific literacy has remained an important and growing topic in science education research, with increasing attention to the development of students' scientific literacy through various learning approaches. Socioscientific issues in science education support scientific literacy by helping prepare students to become informed citizens, while also developing reasoning, argumentation, and understanding of science in social contexts (Eidin & Shwartz, 2023).

Nefianthi & Dwinanda Wahab (2024) found that junior high school students' scientific literacy achievement varied across the PISA-based competencies, with 64.18% for explaining phenomena scientifically and 43.28% for evaluating and designing scientific investigations. The study concluded that students' scientific literacy on Earth topics fell into the moderate category, suggesting the development of scientific literacy skills through levels of inquiry as science practice. Meanwhile, a study of teacher-made tests in Indonesia found that the proportion of higher-order thinking skills (HOTS) questions was lower than that of lower-order

thinking skills (LOTS), indicating that teacher-made assessments still tended to emphasize lower cognitive levels (Apriana et al., 2022). Since scientific literacy aspects are categorized under higher-order thinking skills, it can be concluded that existing teacher-made questions have not properly measured students' scientific literacy. Given this situation and as an effort to improve the quality of science learning assessment instruments—specifically to measure students' scientific literacy—development research is needed to produce scientific literacy assessment instruments that meet standards based on the PISA framework.

Developing a scientific literacy assessment instrument by referencing the PISA framework can serve as a breakthrough in developing junior high school science test items, which currently do not fully measure scientific literacy. Thus, junior high school students are expected to develop reasoning abilities to compete nationally and internationally. Atikah et al. (2022) A good-quality test instrument is essential for accurately measuring students' abilities and providing reliable information about their learning achievement. Written test instruments, beyond determining student capability profiles, can also serve as tools to train student reasoning abilities. Practice questions can include prompts that test student reasoning, including scientific literacy.

Based on the issues outlined above, given the interconnection between assessment and learning, and considering the importance of training and assessing scientific literacy, there is a need to develop assessment instruments capable of measuring junior high school students' scientific literacy based on the PISA framework. This can serve as a guide for science teachers in delivering school science materials equipped to foster scientific literacy and utilize standardized scientific literacy assessment instruments.

Literature Review

In essence, assessment is the process of collecting and processing information to measure student learning achievements and demonstrate learning progress. Assessment is an inseparable process in learning. Educational assessment involves gathering evidence about students' learning to support judgments about their learning and inform instructional and learning decisions (Harris et al., 2022). Assessment can be conducted using various approaches and methods to obtain information on student learning outcomes or competence attainment.

Assessment can be approached through three perspectives: assessment of learning, assessment for learning, and assessment as (Schellekens et al., 2021). Assessment of learning is conducted after the learning process is completed. When teachers perform assessments intended to provide recognition of learning achievements after the learning process ends, they are conducting assessment of learning. Summative assessment forms constitute assessment of learning. Assessment for learning occurs during the learning process and is typically used as a basis to improve teaching and learning. Through assessment for learning, teachers provide feedback on student learning processes, monitor progress, and determine learning advancements. Various formative assessments exemplify assessment for learning. Assessment as learning functions similarly to assessment for learning, operating formatively during the learning process. The difference lies in active student involvement; students gain experience learning to evaluate themselves. Self-assessment and peer assessment exemplify assessment as learning. In assessment as learning, learners can also be involved in formulating assessment procedures, criteria, or rubrics/guidelines so they know precisely what must be done to achieve maximum learning outcomes.

Assessment can be conducted using various methods, including selected-response, written-response, performance-based, and personal communication approaches, depending on the learning outcomes being assessed. Selected response tests include four types: multiple

choice, true/false, matching, and short answer fill-in. Assessment targets can include students' knowledge, reasoning, skills, and ability to produce or demonstrate learning outcomes, with the assessment method selected according to the specific competency being measured.

Instrument development is an important process for producing assessment tools that can appropriately measure the intended abilities or learning outcomes (Adi et al., 2022). A test is an instrument presenting tasks as questions or commands for students to complete. A test is a systematic assessment procedure designed to obtain evidence of students' abilities or learning outcomes through their responses to a set of tasks or items. Test instrument development can follow nine steps: (1) developing test specifications, (2) writing test items, (3) reviewing and revising test items, (4) conducting trials, (5) analyzing test items, (6) revising the test, (7) assembling the test, (8) administering the test, and (9) interpreting test results (Alfarisa et al., 2025). Good test instruments enhance evaluation results that reflect test-takers' capability profiles.

Test instrument quality can be determined by analyzing validity, reliability, discrimination power, and difficulty level. Testing individual items is necessary to determine item quality before performing overall test quality analysis. Atikah et al. (2022) Analyzing test quality is an important step in evaluating the quality of a test and its individual items to ensure that the instrument provides accurate information about students' abilities. Test quality is evaluated by analyzing validity, reliability, discrimination power, and facility level.

Test validity concerns the extent to which an assessment instrument appropriately measures the intended construct and provides evidence that the items function as intended (Mulyanti et al., 2022). Test quality can be examined through evidence of validity and reliability, with reliability referring to the consistency of measurement results (Adi et al., 2022). A good assessment instrument should demonstrate adequate validity and reliability so that it can provide dependable evidence about the construct being measured (Yuniarto et al., 2022).

Item discrimination refers to the extent to which an assessment item differentiates between test-takers with different levels of ability (Sweeney et al., 2022). Discrimination index analysis aims to determine how well an item distinguishes test-takers who master the material from those who do not. The classical item difficulty index is based on the proportion of test-takers who answer an item correctly; therefore, a higher proportion of correct responses indicates a relatively easier item (Jari, 2022). Item facility index analysis aims to classify item difficulty as easy, moderate, or difficult.

PISA (Program for International Student Assessment) is an official international assessment development body conducting international studies on students' scientific literacy. Developed by developed nations in the OECD headquartered in Paris, France, it has been conducted every 3 years since 2000. PISA assesses the knowledge and skills of 15-year-old students across three core domains: reading, mathematics, and science, with an emphasis on applying knowledge and skills to relevant problems. PISA assesses the knowledge and skills of 15-year-old students across three core domains - reading, mathematics, and science - with an emphasis on how students apply their knowledge and skills to relevant and real-life problems (Pokropek et al., 2022).

Scientific literacy is defined as the ability to use scientific knowledge (content, procedural, epistemic) to explain natural phenomena scientifically, evaluate and design scientific inquiries, and interpret data and evidence scientifically. Scientific literacy is multidimensional, involving scientific knowledge, scientific competencies, and scientific contexts in which scientific knowledge and skills are applied (Zhang et al., 2023).

A test blueprint (grid) is a matrix format containing item specification criteria. Test blueprint functions include: 1) guiding item writing, 2) producing items aligned with test goals, and 3) generating relatively equivalent test sets regarding depth and content coverage across different writers. A test blueprint should be developed based on the content and competencies to be measured to ensure that the assessment items are aligned with the intended learning outcomes (Adi et al., 2022). Therefore, scientific literacy assessment blueprints are developed based on junior high science content and measured scientific literacy aspects. Content selection relies on curriculum content standards. The test blueprint describes competencies, content, and scientific literacy aspects to be tested. The blueprint's purpose is to specify exact content scopes and scientific literacy aspects to guide item writing. Scientific literacy assessment based on the PISA framework integrates scientific context, scientific knowledge, and scientific competencies (Zhang et al., 2023).

To achieve national education goals and face 21st-century challenges characterized by rapid technological advances, natural science serves as an essential foundation for national development. Consequently, natural science learning is expected to support students in developing 21st-century skills, including learning and innovation skills such as creativity and innovation, critical thinking and problem-solving, communication, and collaboration; information, media, and ICT literacy; and life and career skills such as flexibility and adaptability, initiative and self-direction, social and cross-cultural skills, productivity and accountability, and leadership and responsibility (Prastiwi et al., 2024). Considering global contexts and Indonesian societal diversity, the 2013 Curriculum's mission is translated into educational practices to equip students with competencies necessary for current and future life. Science learning is expected to develop students' scientific knowledge, scientific process skills, and attitudes toward science, enabling them to engage in scientific activities and apply scientific understanding in relevant contexts (Ernawati et al., 2021).

Science is a systematic effort to create, build, and organize knowledge about natural phenomena. Science serves as a platform to continuously cultivate student attitudes, knowledge, and skills across educational levels. Science learning activities can be implemented through a scientific approach consisting of observing, questioning, experimenting, reasoning, and communicating (Dewi Suprihatin, 2021). Formulating Basic Competencies for Junior High School Science involves overall science competencies and targeted learning outcomes. These competencies are:

1. Living life with positive, honest, and open attitudes; critical, creative, and innovative thinking; and collaboration based on the nature of science.
2. Understanding surrounding natural phenomena based on science learning across Physics, Chemistry, and Biology.
3. Evaluating societal thoughts based on scientific principles and ethics.
4. Solving problems and making decisions in life based on scientific principles and ethics.
5. Identifying and contributing to solving human problems, such as food security, health, energy crises, and the environment.
6. Understanding the impacts of science development on technology and human life - past, present, and potential future impacts - for oneself, others, and the environment.

The 2013 Curriculum requires authentic and non-authentic assessments. In authentic assessments, students apply learned knowledge and skills to solve real-world problems. Authentic assessment (Sukma et al., 2022):

1. Real life application students demonstrate and apply their knowledge, skills, and attitudes in meaningful and real-world contexts.

2. Holistic assessment students' competencies are evaluated comprehensively by considering knowledge, skills, and attitudes directly and holistically evaluates attitudes, knowledge, and skills based on real-world applications.
3. Performance-based demonstration students demonstrate their competencies through meaningful learning tasks rather than relying solely on selected-response tests. uses diverse forms and techniques (performance, products, projects, portfolios, etc.) aligned with measured competencies.
4. Diverse assessment techniques authentic assessment can involve performance assessment, projects, portfolios, and product assessment.
5. Competency-oriented assessment assessment techniques are selected according to the competencies and learning outcomes being measured

Core competencies in the 2013 Curriculum align with achieving scientific literacy, which has become a goal of Indonesia's science curriculum.

2. METHODS

This study utilized a Research and Development (R & D) design. This method is used because the research focused on developing test instruments to measure scientific literacy in physics content. The development process followed the 4D model, consisting of four stages: define, design, develop, and disseminate (Fattikasari & Nasrudin, 2023).

The study was located in Bandung City. Research subjects were public junior high school students in Bandung participating in limited and main trials. Purposive sampling was employed to select samples that had studied the tested material. The sample was taken from a junior high school in the city of Bandung. For analysis purposes, subject schools were categorized into three zoning-based groups to observe scientific literacy patterns.

Data collection techniques included test and non-test methods. Test methods gathered quantitative data, while non-test methods gathered qualitative data via evaluation expert reviews. Data collection was conducted using a scientific literacy assessment instrument that had been previously developed and tested.

3. RESULTS AND DISCUSSION

Expert validation aimed to assess item feasibility prior to research use. Three validators reviewed the instrument, which comprised 30 multiple-choice items accompanied by simulation stimuli. Expert judgment was conducted using blueprints and item indicator rubrics.

Initially, validators assessed item relevance against multiple-choice writing guidelines, reviewing content quality, construct, and language. Ratings used criteria (Unsuitable, Less Suitable, Moderately Suitable, Suitable, Very Suitable) across these three guidelines. Test instrument quality was evaluated across content quality, construct, and language. Results are shown in **Figure 1**.

	Count	Mean	S.D.
Responses non-extreme estimable	= 117	4.71	0.54
Responses in one extreme score	= 153	5.00	0.00
All Responses	= 270	4.87	0.38
Count of measurable responses	= 270		
Raw-score variance of observations	= 0.292	100.00%	
Variance explained by Rasch measures	= 0.050	20.03%	
Variance of residuals	= 0.242	79.97%	

Figure 1. Unidimensionality Results for Scientific Literacy Test Instrument Expert Validation

From **Figure 1**, unidimensionality is indicated by the variance explained by Rasch measures at 20.03%. Based on these results, the test instrument demonstrates adequate quality for measuring students' scientific literacy skills across the assessed (Yuniarto et al., 2022).

Furthermore, item validity was evaluated using Outfit MNSQ, Outfit ZSTD, and Pt-Measure Correlation to determine item fit within the Rasch measurement model (Suherman & Vidákovich, 2025). These values for expert validation data are shown in **Figure 2**.

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Outfit MnSq	Outfit ZStd	Estim. (Discrm)	Correlation PtMea	PtExp	N Aspek
437	90	4.86	4.73	.28	.31	1.07	.3	1.22	.7	.50	.47	1 Kualitas Konten
445	90	4.94	4.90	-.82	.46	.67	-.6	.37	-1.2	.48	.32	2 Konstruksi
434	90	4.82	4.66	.55	.29	1.18	.7	1.12	.4	.53	.51	3 Tata Bahasa
438.7	90.0	4.87	4.76	.00	.35	.97	.1	.90	.0	.50		Mean (Count: 3)
4.6	.0	.05	.10	.59	.08	.22	.6	.38	.9	.02		S.D. (Population)
5.7	.0	.06	.13	.73	.10	.27	.8	.46	1.0	.02		S.D. (Sample)

Model, Populn: RMSE .36 Adj (True) S.D. .47 Separation 1.29 Strata 2.06 Reliability .63
 Model, Sample: RMSE .36 Adj (True) S.D. .63 Separation 1.73 Strata 2.64 Reliability .75
 Model, Fixed (all same) chi-squared: 6.4 d.f.: 2 significance (probability): .04
 Model, Random (normal) chi-squared: 1.6 d.f.: 1 significance (probability): .21

Figure 2. Outfit MnSq, ZStd, and PT-Measure Correlation Results

To determine item validity across quality aspects, values were interpreted against Rasch model fit criteria. Based on expert validation, the scientific literacy instrument was declared valid for 1) content quality, 2) construct, and 3) language.

Expert validation also produced inter-rater reliability values to evaluate validator consensus. Results are shown in **Figure 3**.

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Outfit MnSq	Outfit ZStd	Estim. (Discrm)	Correlation PtMea	PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N Validator
441	90	4.90	4.84	2.27	.36	.91	-.1	.85	-.1	.47	.41	85.6	85.7	1 A
439	90	4.88	4.79	2.03	.34	1.17	.6	.96	.0	.48	.45	86.7	85.4	2 B
436	90	4.84	4.73	1.72	.31	1.05	.2	.91	-.1	.53	.49	83.3	84.7	3 C
438.7	90.0	4.87	4.79	2.01	.34	1.04	.2	.90	-.1	.49				Mean (Count: 3)
2.1	.0	.02	.05	.23	.02	.11	.3	.05	.1	.02				S.D. (Population)
2.5	.0	.03	.06	.28	.03	.13	.4	.06	.1	.03				S.D. (Sample)

Model, Populn: RMSE .34 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability (not inter-rater): .00
 Model, Sample: RMSE .34 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability (not inter-rater): .00
 Model, Fixed (all same) chi-squared: 1.4 d.f.: 2 significance (probability): .50
 Model, Random (normal) chi-squared: .9 d.f.: 1 significance (probability): .35
 Inter-Rater agreement opportunities: 270 Exact agreements: 230 = 85.2% Expected: 230.2 = 85.2%

Figure 3. Inter-Rater Reliability Results for Expert Validation Data

In logical validation, the validator person reliability was 0.69 (moderately reliable). Inter-rater agreement showed exact agreements at 85.2% and expected agreements at 85.2%. Observed agreement closely matched expected Rasch model agreement, indicating consistent ratings across validators.

Reliability relates to measurement stability and consistency (Sitanggang et al., 2025). A test is considered highly reliable when it produces consistent and stable measurement results (Arbeni et al., 2025). Reliability values are shown in **Figure 4**.

Model, Populn: RMSE .36 Adj (True) S.D. .47 Separation 1.29 Strata 2.06 Reliability .63
Model, Sample: RMSE .36 Adj (True) S.D. .63 Separation 1.73 Strata 2.64 Reliability .75
Model, Fixed (all same) chi-squared: 6.4 d.f.: 2 significance (probability): .04
Model, Random (normal) chi-squared: 1.6 d.f.: 1 significance (probability): .21

Figure 4. Item Reliability from Expert Validation

Based on **Figure 4**, the test instrument achieved a reliability value of 0.63, categorized as good. Revised test instruments were tried out on 30 Grade 9 students at a public junior high school in Bandung. Students answered via Google Forms containing items and simulation re-try options. Students also completed readability questionnaires rating item readability,

answer clarity, and comprehension ease/difficulty. Google Forms evaluated item writing rules and readability justifications.

Unidimensionality indicates whether the items in an instrument measure a single underlying construct or latent trait (Mallinson et al., 2022). Unidimensionality values appear in Figure 5.

Responses non-extreme estimable	=	Count	Mean	S.D.
Responses in one extreme score	=	840	3.92	0.97
All Responses	=	60	5.00	0.00
Count of measurable responses	=	900	3.99	0.98
Raw-score variance of observations	=	900		
Variance explained by Rasch measures	=	0.948	100.00%	
Variance of residuals	=	0.571	60.24%	
	=	0.377	39.76%	

Figure 5. Unidimensionality Value

From Figure 5, unidimensionality (variance explained by Rasch measures) reached 60.24%, categorizing the instrument as sufficient to measure student readability. Inter-rater reliability evaluated student consensus on readability. Readability inter-rater results appear in Figure 6.

Table 7.1.3 Siswa Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Infit S.E.	Outfit MnSq ZStd	Estim. (Discrm)	Correlation PtMea PtExp	Exact Agree. Obs % Exp %	Nu Siswa
136	30	4.53	4.55	3.53	.32	2.19 3.6	1.84 2.6	-.10 .37	33.3 35.3	1 1
105	30	3.50	3.50	1.14	.26	1.57 2.0	1.57 2.0	.35 .53	28.5 27.3	2 2
148	30	4.93	4.94	5.86	.73	.95 .1	.74 -.1	1.04 .22	36.2 35.9	3 3
150	30	5.00	4.99	(7.81 1.83)Maximum				.00 .00	35.7 35.7	4 4
96	30	3.20	3.20	.53	.26	1.10 .4	1.10 .4	.90 .23	23.6 23.6	5 5
116	30	3.87	3.87	1.90	.27	.83 -.6	.85 -.5	1.17 -.05	30.3 30.9	6 6
62	30	2.07	2.06	-1.99	.29	.73 -1.0	.67 -1.3	1.39 .35	8.0 7.7	7 7
120	30	4.00	4.01	2.18	.27	.34 -3.6	.36 -2.5	1.78 .20	30.8 32.0	8 8
126	30	4.20	4.21	2.64	.28	.55 -2.2	.57 -2.0	1.52 .21	32.0 32.5	9 9
90	30	3.00	3.00	.12	.26	.14 -5.5	.14 -5.6	1.88 .00	19.8 20.9	10 10
89	30	2.97	2.96	.05	.26	.23 -4.5	.23 -4.4	1.79 .34	19.4 20.5	11 11
116	30	3.87	3.87	1.90	.27	.62 -1.7	.62 -1.7	1.42 .07	29.1 30.9	12 12
115	30	3.83	3.84	1.82	.26	1.95 3.1	1.99 3.3	-.23 .01	36.6 30.6	13 13
150	30	5.00	4.99	(7.81 1.83)Maximum				.00 .00	35.7 35.7	14 14
148	30	4.93	4.94	5.86	.73	1.01 .2	1.16 .4	.97 -.01	36.0 35.9	15 15
146	30	4.87	4.88	5.11	.53	.92 .0	.77 -.3	1.05 .26	35.9 36.0	16 16
118	30	2.93	2.94	2.04	.27	1.85 2.9	1.80 2.8	-.06 .51	32.3 31.5	17 17
124	30	4.13	4.15	2.48	.28	3.58 6.7	3.27 6.4	-2.30 .44	30.8 33.1	18 18
149	30	4.97	4.97	6.58	1.01	.97 .2	.66 .0	1.03 .18	36.0 35.8	19 19
108	30	3.60	3.60	1.34	.26	.51 -2.3	.51 -2.3	1.55 .43	27.6 28.4	20 20
132	30	4.40	4.42	3.14	.30	1.54 2.0	1.45 1.7	.45 .37	36.6 34.7	21 21
125	30	4.17	4.18	2.56	.28	.44 -2.9	.45 -2.8	1.70 .36	33.4 32.3	22 22
102	30	3.43	3.43	1.00	.26	.77 -.9	.77 -.9	1.25 .38	25.5 26.5	23 23
106	30	3.53	3.53	1.21	.26	.76 -.9	.76 -.9	1.29 .69	27.6 28.5	24 24
140	30	4.67	4.69	4.00	.36	1.75 2.3	1.94 2.4	.44 .03	36.3 35.8	25 25
124	30	4.13	4.15	2.48	.28	.60 -1.8	.64 -1.6	1.52 .44	33.9 33.1	26 26
81	30	2.70	2.69	-.51	.27	.94 -.1	.92 -.2	1.12 .74	14.3 16.7	27 27
149	30	4.97	4.97	6.58	1.01	1.03 .3	1.83 .9	.93 -.17	35.4 35.8	28 28
115	30	3.83	3.84	1.82	.26	.29 -4.1	.28 -4.2	1.83 .35	36.1 30.6	29 29
103	30	3.43	3.43	1.00	.26	.67 -1.4	.67 -1.4	1.37 .52	28.3 26.5	30 30
119.7	30.0	3.99	3.99	2.73	.47	1.03 -.4	1.02 -.4	.27		Mean (Count: 30)
22.5	.0	.75	.75	2.43	.42	.72 2.7	.70 2.6	.22		S.D. (Population)
22.9	.0	.76	.77	2.47	.43	.73 2.7	.72 2.7	.23		S.D. (Sample)
With extremes, Model, Populn: RMSE .63 Adj (True) S.D. 2.35 Separation 3.73 Strata 5.31 Reliability (not inter-rater) .93										
With extremes, Model, Sample: RMSE .63 Adj (True) S.D. 2.39 Separation 3.80 Strata 5.40 Reliability (not inter-rater) .94										
Without extremes, Model, Populn: RMSE .43 Adj (True) S.D. 2.04 Separation 4.76 Strata 6.68 Reliability (not inter-rater) .96										
Without extremes, Model, Sample: RMSE .43 Adj (True) S.D. 2.08 Separation 4.85 Strata 6.80 Reliability (not inter-rater) .96										
With extremes, Model, Fixed (all same) chi-squared: 659.8 d.f.: 29 significance (probability): .00										
With extremes, Model, Random (normal) chi-squared: 25.3 d.f.: 28 significance (probability): .96										
Inter-Rater agreement opportunities: 13050 Exact agreements: 3873 = 29.7% Expected: 3524.7 = 30.1%										

Figure 6. Inter-Rater Reliability Results for Test Readability

For test readability, student person reliability was 0.93 (sufficient reliability). Exact agreement reached 29.7% versus 30.1% expected agreement, showing student consensus aligned with Rasch model expectations. Readability percentage results appear in Figure 7.

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Infit S.E.	Outfit MnSq ZStd	Estim. (Discrm)	Correlation PtMea PtExp	N Aspek
3590	900	3.99	4.10	.00	.06	1.00 .0	1.02 .3	1.00 .75	1 Keterbacaan

Figure 7. Inter-Rater Reliability Results for Test Readability

Results show the instrument generally exhibited clear readability. Students noted that phrasing was straightforward with good grammar.

In the main trial stage, 9th-grade public junior high school students in Bandung participated to evaluate coherent knowledge. Test validity was analyzed using Many Facet Rasch Measurement (MFRM) via Minifac software. Trial results evaluated unidimensionality, outfit mean square (MNSQ), outfit Z-standard (ZSTD), and Point Measure Correlation (Pt Mean Corr) to verify item fit criteria. Unidimensionality values appear in **Figure 8**.

```

Responses non-extreme estimable = 2000 0.43 0.50
Count of measurable responses = 2000
Raw-score variance of observations = 0.245 100.00%
Variance explained by Rasch measures = 0.053 21.44%
Variance of residuals = 0.193 78.56%

```

Figure 8. Unidimensionality Results

Figure 8 displays variance explained by Rasch measures at 21.44%, categorizing the test instrument as sufficient. Outfit MnSq, ZStd, and PT-Measure Correlation item-level fit values are shown in **Figure 9**.

Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrim	Correlation PtMea	PtExp	Nu items
1.23	.32	1.11	.6	1.19	.7	.85	.07	.25	1 1
-.69	.26	.93	-.8	.90	-.9	1.34	.42	.31	2 2
-1.19	.28	.86	-1.2	.77	-1.6	1.39	.51	.30	3 3
-.03	.26	1.02	.2	1.05	.5	.88	.27	.31	4 4
-1.35	.28	.86	-1.0	.80	-1.1	1.30	.49	.30	5 5
.18	.26	1.17	1.8	1.32	2.5	.25	.04	.30	6 6
.46	.27	1.03	.3	1.08	.5	.88	.23	.29	7 7
-2.36	.37	.87	-.4	.74	-.7	1.12	.43	.24	8 8
1.13	.31	.93	-.3	.87	-.4	1.10	.35	.25	9 9
-.09	.26	.90	-1.3	.88	-1.2	1.52	.45	.31	10 10
-1.19	.28	.86	-1.2	.79	-1.4	1.37	.51	.30	11 11
.04	.26	1.03	.4	1.04	.3	.84	.26	.31	12 12
-.16	.26	.92	-1.1	.88	-1.3	1.49	.44	.31	13 13
.46	.27	1.01	.1	.98	.0	.98	.28	.29	14 14
-.09	.26	.84	-2.2	.81	-2.0	1.84	.53	.31	15 15
-1.04	.27	.86	-1.3	.85	-1.1	1.39	.48	.31	16 16
1.33	.33	1.12	.6	1.20	.7	.85	.06	.24	17 17
-1.69	.31	1.00	.0	.88	-.5	1.04	.31	.28	18 18
.61	.28	1.19	1.5	1.22	1.2	.57	.02	.29	19 19
-.03	.26	1.15	1.9	1.16	1.4	.27	.10	.31	20 20
.23	.27	1.17	1.8	1.19	1.4	.35	.06	.30	21 21
-.38	.26	1.02	.2	1.00	.0	.92	.29	.31	22 22
1.12	.31	1.09	.5	1.33	1.3	.82	.07	.25	23 23
.52	.28	1.06	.5	1.06	.4	.84	.20	.29	24 24
.93	.30	.94	-.3	1.28	1.3	1.04	.29	.27	25 25
1.02	.30	1.10	.6	1.14	.6	.84	.11	.26	26 26
-.38	.26	.90	-1.3	.87	-1.4	1.56	.45	.31	27 27
.16	.26	.88	-1.4	.85	-1.2	1.51	.47	.30	28 28
.23	.27	1.06	.6	1.11	.9	.73	.20	.30	29 29
1.02	.30	1.04	.3	1.23	1.0	.89	.16	.26	30 30

.00	.28	1.00	-.1	1.02	.0		.28	Mean (Count: 30)	
.91	.03	.11	1.1	.18	1.2		.17	S.D. (Population)	
.93	.03	.11	1.1	.18	1.2		.17	S.D. (Sample)	

Figure 9. Outfit MnSq, ZStd, and PT-Measure Correlation Results

Output analysis confirmed that all items fit, demonstrating validity for research use. Measurement reliability concerns the stability and consistency of test scores across repeated administrations, including variations in time, raters, and administration condition (De Vet et al., 2017) Reliability results appear in **Figure 10**.

```

Model, Populn: RMSE .28 Adj (True) S.D. .87 Separation 3.07 Strata 4.43 Reliability .90
Model, Sample: RMSE .28 Adj (True) S.D. .88 Separation 3.13 Strata 4.50 Reliability .91
Model, Fixed (all same) chi-squared: 264.4 d.f.: 29 significance (probability): .00
Model, Random (normal) chi-squared: 26.2 d.f.: 29 significance (probability): .56

```

Figure 10. Test Trial Reliability

Figure 10 shows a reliability value of 0.90, categorized as very good. Item discrimination power differentiates high- and low-ability students. PT-Measure Correlation values are displayed in **Figure 11**.

-	Model	Infit	Outfit	Estim.	Correlation				
Measure	S.E.	MnSq	ZStd	MnSq	ZStd	Discrm	PtMea	PtExp	Nu items
1.23	.32	1.11	.6	1.19	.7	.85	.07	.25	1 1
-.69	.26	.93	-.8	.90	-.9	1.34	.42	.31	2 2
-1.19	.28	.86	-1.2	.77	-1.6	1.39	.51	.30	3 3
-.03	.26	1.02	.2	1.05	.5	.88	.27	.31	4 4
-1.35	.28	.86	-1.0	.80	-1.1	1.30	.49	.30	5 5
.18	.26	1.17	1.8	1.32	2.5	.25	.04	.30	6 6
.46	.27	1.03	.3	1.08	.5	.88	.23	.29	7 7
-2.36	.37	.87	-.4	.74	-.7	1.12	.43	.24	8 8
1.13	.31	.93	-.3	.87	-.4	1.10	.35	.25	9 9
-.09	.26	.90	-1.3	.88	-1.2	1.52	.45	.31	10 10
-1.19	.28	.86	-1.2	.79	-1.4	1.37	.51	.30	11 11
.04	.26	1.03	.4	1.04	.3	.84	.26	.31	12 12
-.16	.26	.92	-1.1	.88	-1.3	1.49	.44	.31	13 13
.46	.27	1.01	.1	.98	.0	.98	.28	.29	14 14
-.09	.26	.84	-2.2	.81	-2.0	1.84	.53	.31	15 15
-1.04	.27	.86	-1.3	.85	-1.1	1.39	.48	.31	16 16
1.33	.33	1.12	.6	1.20	.7	.85	.06	.24	17 17
-1.69	.31	1.00	.0	.88	-.5	1.04	.31	.28	18 18
.61	.28	1.19	1.5	1.22	1.2	.57	.02	.29	19 19
-.03	.26	1.15	1.9	1.16	1.4	.27	.10	.31	20 20
.23	.27	1.17	1.8	1.19	1.4	.35	.06	.30	21 21
-.38	.26	1.02	.2	1.00	.0	.92	.29	.31	22 22
1.12	.31	1.09	.5	1.33	1.3	.82	.07	.25	23 23
.52	.28	1.06	.5	1.06	.4	.84	.20	.29	24 24
.93	.30	.94	-.3	1.28	1.3	1.04	.29	.27	25 25
1.02	.30	1.10	.6	1.14	.6	.84	.11	.26	26 26
-.38	.26	.90	-1.3	.87	-1.4	1.56	.45	.31	27 27
.16	.26	.88	-1.4	.85	-1.2	1.51	.47	.30	28 28
.23	.27	1.06	.6	1.11	.9	.73	.20	.30	29 29
1.02	.30	1.04	.3	1.23	1.0	.89	.16	.26	30 30
.00	.28	1.00	-.1	1.02	.0		.28		Mean (Count: 30)
.91	.03	.11	1.1	.18	1.2		.17		S.D. (Population)
.93	.03	.11	1.1	.18	1.2		.17		S.D. (Sample)

Figure 11. Outfit MnSq, ZStd, and PT-Measure Correlation Values

PT-Measure Correlation indicates discrimination power. Items 10, 9, and 19 showed negative correlation values, indicating misleading items where low-ability students answered correctly while high-ability students answered incorrectly. Negative items must be checked for answer key errors, revised, or removed. Items 17, 11, 13, 7, 14, 16, 6, 2, 4, 21, and 3 required revision due to poor discrimination capacity. Items 8, 12, and 18 fell into the good category, while items 15, 20, and 1 performed exceptionally well in distinguishing high- and low-ability students.

Difficulty level indicates how easy or difficult an item is. In Rasch modeling, item difficulty (measure logit) reflects log-transformed odd-ratios. Higher logit values signify higher difficulty. Difficulty values are shown in **Figure 12**.

Total	Total	Obsvd	Fair(H)	-	Model	Infit	Outfit	Estim.	Correlation				
Scores	Count	Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	Discrm	PtMea	PtExp	Nu items
12	67	.18	.18	1.33	.32	1.11	.6	1.19	.7	.85	.07	.25	17 17
19	67	.17	.17	1.23	.32	1.11	.6	1.19	.7	.85	.07	.25	1 1
14	67	.21	.19	1.13	.31	.93	-.3	.97	-.4	1.10	.35	.25	9 9
16	66	.21	.19	1.12	.31	1.05	.8	1.33	1.3	.82	.07	.25	23 23
15	66	.23	.21	1.02	.30	1.10	.6	1.14	.6	.94	.11	.26	26 26
18	66	.23	.21	1.02	.30	1.04	.3	1.23	1.0	.89	.16	.26	30 30
16	66	.24	.22	.99	.30	.94	-.3	1.28	1.3	1.04	.29	.27	25 25
20	67	.30	.28	.61	.28	1.19	1.5	1.22	1.2	.57	.02	.29	19 19
21	66	.32	.30	.82	.28	1.06	.5	1.06	.4	.84	.20	.29	24 24
22	67	.33	.31	.46	.27	1.03	.3	1.08	.5	.88	.27	.31	7 7
22	67	.33	.31	.46	.27	1.01	.1	.98	.0	.98	.28	.29	14 14
35	66	.38	.36	.23	.27	1.17	1.9	1.19	1.4	.35	.06	.30	21 21
25	66	.38	.36	.23	.27	1.04	.6	1.11	.9	.73	.20	.30	29 29
26	67	.39	.38	.18	.26	1.17	1.8	1.32	2.5	.25	.04	.30	6 6
36	66	.39	.38	.18	.26	.88	1.4	.86	-1.2	1.51	.47	.30	28 28
28	67	.42	.41	.04	.26	1.03	.4	1.04	.3	.94	.26	.31	12 12
29	67	.43	.43	-.03	.26	1.02	.3	1.05	.8	.88	.27	.31	4 4
29	67	.43	.43	-.03	.26	1.15	1.9	1.16	1.4	.37	.10	.31	20 20
30	67	.44	.44	-.09	.26	.90	-1.3	.88	-1.2	1.52	.45	.31	10 10
30	67	.45	.44	-.09	.26	.84	-2.3	.81	-2.0	1.84	.53	.31	15 15
31	67	.46	.46	-.14	.26	.92	-1.1	.98	-1.2	1.49	.44	.31	13 13
34	66	.53	.51	-.38	.26	1.02	.3	1.00	.0	.90	.25	.31	22 22
34	66	.52	.51	-.39	.26	.90	-1.9	.97	-1.4	1.56	.45	.31	27 27
35	67	.55	.53	-.83	.26	.93	-.8	.90	-.3	1.34	.42	.31	2 2
46	67	.66	.67	-1.04	.27	.86	-1.3	.86	-1.1	1.35	.49	.31	16 16
46	67	.65	.70	-1.19	.28	.86	-1.2	.77	-1.6	1.39	.48	.31	3 3
46	67	.65	.70	-1.19	.28	.86	-1.2	.79	-1.4	1.37	.51	.30	11 11
48	67	.72	.73	-1.25	.28	.86	-1.0	.89	-1.1	1.30	.49	.30	5 5
52	67	.70	.80	-1.63	.31	1.00	.0	.88	-.5	1.04	.31	.28	18 18
58	67	.87	.88	-3.46	.37	.87	-.4	.74	-.7	1.12	.43	.24	8 8
29.5	66.7	.43	.42	-.50	.28	1.00	-.1	1.02	.0		.28		Mean (Count: 30)
12.8	.5	.18	.20	-.91	.03	.11	1.1	.18	1.2		.17		S.D. (Population)
12.5	.5	.19	.20	-.93	.03	.11	1.1	.18	1.2		.17		S.D. (Sample)

Figure 12. Item Difficulty Values

Figure 12 displays logit estimates. Item 7 (2.30 logit) was the most difficult item (14 correct responses). Item 1 (-2.68 logit) was the easiest item (138 correct responses). Testing 150 students involved gender dynamics, which can introduce measurement bias if items favor one gender over another. Gender bias analysis results appear in **Figure 13**.

Bias+	Model				Infit	Outfit	gender		items			
Size	S.E.	t	d.f.	Prob.	MnSq	MnSq	Sq	N	gender	measr-	Nu	it measr-
-.25	.45	-.56	38	.5792	1.1	1.1	1	1	Perempuan	.02	1	1.23
.29	.46	.64	27	.5258	1.2	1.3	2	2	Laki-laki	-.02	1	1.23
-.29	.33	-.86	38	.3952	1.0	1.0	3	1	Perempuan	.02	2	-.69
.45	.43	1.04	27	.3063	.9	.8	4	2	Laki-laki	-.02	2	-.69
.17	.37	.47	38	.6433	.9	.8	5	1	Perempuan	.02	3	-1.19
-.23	.42	-.55	27	.5864	.8	.8	6	2	Laki-laki	-.02	3	-1.19
.15	.34	.46	38	.6504	1.1	1.1	7	1	Perempuan	.02	4	-.03
-.22	.41	-.55	27	.5901	.9	.9	8	2	Laki-laki	-.02	4	-.03
-.49	.34	-1.42	38	.1642	.8	.8	9	1	Perempuan	.02	5	-1.35
.96	.57	1.69	27	.1026	.8	.7	10	2	Laki-laki	-.02	5	-1.35
.24	.34	.72	38	.4729	1.2	1.2	11	1	Perempuan	.02	6	.18
-.37	.43	-.86	27	.3954	1.2	1.7	12	2	Laki-laki	-.02	6	.18
.64	.34	1.92	38	.0629	1.0	1.0	13	1	Perempuan	.02	7	.46
-1.22	.56	-2.19	27	.0371	.9	.9	14	2	Laki-laki	-.02	7	.46
.05	.49	.10	38	.9230	.9	.7	15	1	Perempuan	.02	8	-2.36
-.06	.57	-.10	27	.9201	.8	.8	16	2	Laki-laki	-.02	8	-2.36
-.16	.43	-.37	38	.7126	1.0	.9	17	1	Perempuan	.02	9	1.13
.19	.46	.43	27	.6737	.9	.9	18	2	Laki-laki	-.02	9	1.13
.09	.34	.26	38	.7982	.8	.8	19	1	Perempuan	.02	10	-.09
-.12	.41	-.31	27	.7616	1.0	1.0	20	2	Laki-laki	-.02	10	-.09
.17	.37	.47	38	.6433	.9	.8	21	1	Perempuan	.02	11	-1.19
-.23	.42	-.55	27	.5864	.8	.7	22	2	Laki-laki	-.02	11	-1.19
.22	.34	.66	38	.5147	1.1	1.1	23	1	Perempuan	.02	12	.04
-.33	.42	-.78	27	.4395	1.0	1.0	24	2	Laki-laki	-.02	12	.04
.13	.33	.39	38	.6954	1.0	1.0	25	1	Perempuan	.02	13	-.16
-.19	.41	-.47	27	.6419	.8	.7	26	2	Laki-laki	-.02	13	-.16
.06	.35	.16	38	.8719	1.1	1.1	27	1	Perempuan	.02	14	.46
-.08	.43	-.20	27	.8464	.8	.8	28	2	Laki-laki	-.02	14	.46
-.38	.35	-1.08	38	.2856	.8	.8	29	1	Perempuan	.02	15	-.09
.53	.41	1.30	27	.2060	.8	.8	30	2	Laki-laki	-.02	15	-.09
-.06	.35	-.18	38	.8589	.8	.8	31	1	Perempuan	.02	16	-1.04
.10	.43	.23	27	.8213	.9	.9	32	2	Laki-laki	-.02	16	-1.04
-1.40	.73	-1.91	38	.0633	1.0	1.2	33	1	Perempuan	.02	17	1.33
.97	.42	2.31	27	.0285	1.2	1.2	34	2	Laki-laki	-.02	17	1.33
-.04	.39	-.10	38	.9206	1.0	.9	35	1	Perempuan	.02	18	-1.69
.07	.49	.14	27	.8927	1.1	.9	36	2	Laki-laki	-.02	18	-1.69
-.34	.39	-.87	38	.3923	1.1	1.3	37	1	Perempuan	.02	19	.61
.42	.41	1.02	27	.3184	1.3	1.2	38	2	Laki-laki	-.02	19	.61
.38	.33	1.13	38	.2667	1.1	1.1	39	1	Perempuan	.02	20	-.03
-.57	.43	-1.34	27	.1910	1.2	1.1	40	2	Laki-laki	-.02	20	-.03
-.30	.36	-.83	38	.4107	1.2	1.2	41	1	Perempuan	.02	21	.23
.41	.41	1.00	26	.3281	1.2	1.2	42	2	Laki-laki	-.02	21	.23
-.31	.34	-.93	38	.3600	.9	.9	43	1	Perempuan	.02	22	-.38
.48	.42	1.13	26	.2689	1.1	1.1	44	2	Laki-laki	-.02	22	-.38
.17	.39	.43	38	.6731	1.2	1.5	45	1	Perempuan	.02	23	1.12
-.26	.51	-.51	26	.6154	1.0	1.0	46	2	Laki-laki	-.02	23	1.12
-.43	.39	-1.09	38	.2820	.9	.9	47	1	Perempuan	.02	24	.52
.53	.41	1.30	26	.2063	1.2	1.2	48	2	Laki-laki	-.02	24	.52
.13	.38	.34	38	.7341	.9	.7	49	1	Perempuan	.02	25	.93
-.20	.48	-.41	26	.6826	1.0	2.3	50	2	Laki-laki	-.02	25	.93
.36	.37	.98	38	.3346	1.0	1.0	51	1	Perempuan	.02	26	1.02
-.64	.56	-1.15	26	.2606	1.2	1.4	52	2	Laki-laki	-.02	26	1.02
.36	.34	1.07	38	.2903	.9	.9	53	1	Perempuan	.02	27	-.38
-.54	.41	-1.31	26	.2024	.9	.8	54	2	Laki-laki	-.02	27	-.38
.23	.34	.68	38	.4998	.9	.9	55	1	Perempuan	.02	28	.16
-.36	.43	-.83	26	.4153	.9	.8	56	2	Laki-laki	-.02	28	.16
-.05	.35	-.14	38	.8865	1.1	1.2	57	1	Perempuan	.02	29	.23
.07	.41	.17	26	.8682	1.0	1.1	58	2	Laki-laki	-.02	29	.23
-.09	.41	-.22	38	.8302	1.1	1.5	59	1	Perempuan	.02	30	1.02
.11	.46	.25	26	.8081	1.0	.9	60	2	Laki-laki	-.02	30	1.02
-.02	.41	.00			1.0	1.0	Mean (Count: 60)					
.42	.07	.93			.1	.3	S.D. (Population)					
.43	.07	.93			.1	.3	S.D. (Sample)					

Figure 13. Test Instrument Bias Analysis

Items exhibit gender bias when probability logit values fall below 0.05 (5%) (Evita et al., 2023). In **Figure 13**, item 10 exhibited gender bias favoring female students due to higher female probability values.

4. CONCLUSION

The developed scientific literacy assessment instrument measures cognitive abilities via 30 multiple-choice items with five options. Items integrate scientific practices, disciplinary core ideas, and crosscutting concepts, utilizing text, images, and graphs as stimuli. All 30 developed test items met validity criteria.

The simulation-based wave and light test using a 3D learning framework met logical validity and empirical validity/reliability based on main trial results. Thus, the instrument effectively measures cognitive abilities in light waves within a 3D framework.

Student readability percentage results showed high clarity ratings for asking questions within scientific practices, as students understood concepts and identified scientific questions easily. Conversely, evaluating information received the highest unclear responses due to item difficulty. Instrument characteristics analyzed via Many Facet Rasch Measurement (MFRM) using Minifac software demonstrated sufficient unidimensionality.

This study utilized a small sample size; therefore, the results may not provide a comprehensive picture. Future research with a larger sample size is recommended.

5. ACKNOWLEDGMENT

This research was funded through the Assigned Annual Work Plan and Budget of the Institute for Research and Community Service Universitas Pendidikan Indonesia, Fiscal Year 2022 under Rector's Decree Number: 0358/UN40/PT.01.02/2022.

6. REFERENCES

- Adi, N. R. M., Amaruddin, H., Adi, H. M. M., and A'yun, I. L. Q. (2022). Validity and reliability analysis using the rasch model to measure the quality of mathematics test items of vocational high schools. *Journal of research and educational research evaluation*.
- Alfarisa, F., Arisetyawan, A., Maulana, F. R., Lathifah, N. Y., Afriani, S., and Mashudi, M. (2025). Creative thinking in elementary education: Reliable assessment instruments for science and social studies. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 29(1), 13–25.
- AlKaabi, N. A., Al-Maadeed, N., Romanowski, M. H., and Sellami, A. (2024). Drawing lessons from PISA: Qatar's use of PISA results. *PROSPECTS*, 54(1), 221–240.
- Alotaibi, W. H., and Alghamdi, A. K. H. (2025). Saudi secondary school science textbooks' ability to inculcate a PISA-Informed Scientific Identity. *PLOS One*, 20(10).
- Apriana, S., Lubis, A. A., and Sofyan, D. (2022). Higher-order thinking skills on teachers-made tests by english teachers of a senior high school in Bengkulu city. *Journal of English Teaching, Applied Linguistics and Literatures (JETALL)*, 5(2), 137–152.
- Arbeni, W., Windiani, A., Sihotang, D. S. B., Anggraini, N., Wulandari, S., and Nugroho, A. (2025). Test reliability analysis in educational evaluation: A quantitative approach to consistency and validity. *Holistic Science*, 5(1), 59–64.
- Atikah, A., Sudiyatno, S., Rahim, A., and Marlina, M. (2022a). Assessing the item of final assessment mathematics test of junior high school using Rasch model. *Jurnal Elemen*, 8(1), 117–130.

- Atikah, A., Sudiyatno, S., Rahim, A., and Marlina, M. (2022b). Assessing the item of final assessment mathematics test of junior high school using Rasch model. *Jurnal Elemen*, 8(1), 117–130.
- De Vet, H. C. W., Mokkink, L. B., Mosmuller, D. G., and Terwee, C. B. (2017). Spearman–Brown prophecy formula and Cronbach’s alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, 85, 45–49.
- Dewi Suprihatin, E. A. (2021). Scientific approach to promote scientific writing skills using blended learning system. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(13), 762–769.
- Eidin, E., and Shwartz, Y. (2023). From Ideal to Practical—A Design of Teacher Professional Development on Socioscientific Issues. *Sustainability*, 15(14), 11394.
- Ernawati, M. D. W., Asrial, A., Perdana, R., Septi, S. E., and Rahmi, R. (2021). Evaluation of Students’ Attitudes and Science Process Skills toward Middle School Science Subject in Indonesia. *Jurnal Pendidikan Progresif*, 11(2), 258–274.
- Evita, E., Sari, E. D. K., Susieni, E., and Tamam, B. (2023). Analysis of item quality in islamic religious education subjects with rasch modeling. *Conciencia*, 23(2), 147–156.
- Fattikasari, D. W., and Nasrudin, H. (2023). Development of electronic worksheets based on project-based learning assisted by a virtual laboratory to train students’ critical thinking skills. *Jurnal Pijar Mipa*, 18(4), 493–501.
- Harris, L. R., Adie, L., and Wyatt-Smith, C. (2022). Learning progression–based assessments: a systematic review of student and teacher uses. *Review of Educational Research*, 92(6), 996–1040.
- Khery, Y., Sarjan, M., Ahzan, S., and Efendi, I. (2022). Konseptualisasi literasi sains mengacu pada kerangka sains pisa sejak tahun 2000. *Educatoria: Jurnal Ilmiah Ilmu Pendidikan*, 2(4), 200–231.
- Mallinson, T., Kozlowski, A. J., Johnston, M. V., Weaver, J., Terhorst, L., Grampurohit, N., Juengst, S., Ehrlich-Jones, L., Heinemann, A. W., Melvin, J., Sood, P., and Van De Winkel, A. (2022). Rasch reporting guideline for rehabilitation research (RULER): The RULER Statement. *Archives of Physical Medicine and Rehabilitation*, 103(7), 1477–1486.
- Jari, M. (2022). Seeking the real item difficulty: Bias-corrected item difficulty and some consequences in Rasch and IRT modeling. *Behaviormetrika*, 50(1), 121–154.
- Mulyanti, S., Sukmawati, W., and Tarkin, N. E. H. (2022). Development of items in acid-base identification experiments using natural materials: validity test with rasch model analysis. *Phenomenon: Jurnal Pendidikan MIPA*, 12(1), 17–30.
- Nefianthi, R., and Dwinanda Wahab, M. N. N. (2024). The analysis of last year junior high school students’ scientific literacy in biology topics. *BIO-INOVED: Jurnal Biologi-Inovasi Pendidikan*, 6(3), 397.

- Pokropek, A., Marks, G. N., Borgonovi, F., Koc, P., and Greiff, S. (2022a). General or specific abilities? Evidence from 33 countries participating in the PISA assessments. *Intelligence*, 92, 101653.
- Pokropek, A., Marks, G. N., Borgonovi, F., Koc, P., and Greiff, S. (2022b). General or specific abilities? Evidence from 33 countries participating in the PISA assessments. *Intelligence*, 92, 101653.
- Prastiwi, C. H. W., Rukmini, D., Saleh, M., and Astuti, P. (2024). Framework for assessing 21st century skills in EFL project-based learning. *Proceedings of UNNES-TEFLIN National Conference*, 6, 328–337.
- Ramadhan, A. N., Dwiningrum, S. I. A., dan Subhan, B. (2021). Potensi integrasi pembelajaran biologi dengan pembelajaran quran-hadis. *AL QUDES: Jurnal Studi Alquran Dan Hadis*, 5(1), 263.
- Schellekens, L. H., Bok, H. G. J., De Jong, L. H., Van Der Schaaf, M. F., Kremer, W. D. J., and Van Der Vleuten, C. P. M. (2021). A scoping review on the notions of Assessment as Learning (AaL), Assessment for Learning (AfL), and Assessment of Learning (AoL). *Studies in Educational Evaluation*, 71, 101094.
- Sitanggang, Y. J., Simorangkir, T., Sihaloho, A. L., Pakpahan, R. S. M., and Silalahi, D. E. (2025). Analysis of english skill tests for grade vii students based on language assessment principles. *JURNAL PENDIDIKAN BAHASA*, 15(3), 349–360.
- Soeharto, S., and Csapó, B. (2022). Assessing Indonesian student inductive reasoning: Rasch analysis. *Thinking Skills and Creativity*, 46, 101132.
- Suherman, S., and Vidákovich, T. (2025). An assessment of creativity style inventory: A Rasch model. *Journal of Creativity*, 35(3), 100110.
- Sukma, Abbas, A., Nurhayati, Kaharuddin, and Gheisari, A. (2022). Development of authentic assessment in local wisdom-based reading learning. *Education Research International*, 2022, 1–11.
- Sweeney, S. M., Sinharay, S., Johnson, M. S., and Steinhauer, E. W. (2022). An investigation of the nature and consequence of the relationship between irt difficulty and discrimination. *Educational Measurement: Issues and Practice*, 41(4), 50–67.
- Syah, R., Darmawan, D., and Syamsuddin, S. (2023). Development of a scientific attitude measurement instrument for Dharmacariya department students at STABN using the Fishbein and Ajzen model. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 27(2), 226–240.
- Yuniarto, B., Yudha, R. P., and Nurchotimah, A. S. I. (2022). Validity and reliability: Social attitude assessment instruments. *Communications in Humanities and Social Sciences*, 2(1), 1–6.
- Zhang, L., Liu, X., and Feng, H. (2023). Development and validation of an instrument for assessing scientific literacy from junior to senior high school. *Disciplinary and Interdisciplinary Science Education Research*, 5(1), 21.

Zuhrotul Wardi, L., Wasis, W., Widodo, W., Satriawan, M., and Nisa, K. (2024). Trends and development in digital learning research for enhancing scientific literacy in physics learning: literature review and bibliometric analysis. *Jurnal Pendidikan MIPA*, 25(3), 1051–1066.