JPF | Volume 14 | Number 1 | 2026 | 119 - 137

p - ISSN: 2302-8939
e - ISSN: 2527-4015

Jurnal Pendidikan Fisika

https://journal.unismuh.ac.id/index.php/jpf

DOI: 10.26618/n0sjn231

Validity and Reliability of Science Literacy Assessment Instruments for Measuring Science Competencies in the Context of PISA 2025 using the Rasch Model

Mentari Darma Putri[1)*], Dini Fitria [2)], Nurlaini [1)], Nur Jannah Berutu[1)]

[1] *Study Program of Physics Education, Samudra University, Langsa, 24416, Indonesia*
[2] *Study Program of Biology Education, Samudra University, Langsa, 24416, Indonesia*

**Corresponding author: mentari_darmap@unsam.ac.id*

Received: October 16, 2025; Accepted: December 23, 2025; Published: January 20, 2026

Abstract – *Science literacy is a key competency for preparing students to reason scientifically, interpret evidence, and make informed decisions in real-world contexts. Yet Indonesian students' performance in international assessments remains below the OECD average, underscoring the urgent need for assessment tools aligned with global standards. This study aimed to develop and validate a PISA 2025–oriented science literacy assessment instrument to support the measurement of senior high school students' competencies in scientifically explaining phenomena; constructing and evaluating investigation designs; critically interpreting data and evidence; and researching, evaluating, and using scientific information for decision-making and action. Using a research and development approach with the 4-D model, ten context-rich essay items and an analytic scoring rubric (levels 1–4) were produced. Content validity was examined by three experts using Aiken's V, and empirical validation was conducted with 50 Year 11 students from two senior high schools in Langsa City using the Rasch Model (Winstep). The instrument demonstrated strong content validity (Aiken's V averages: construction = 0.84, relevance = 0.89, clarity = 0.90). Rasch results showed that 8 of 10 items met fit criteria, while two items displayed misfit on selected indices but remained positively correlated with the measured construct. Reliability was high (Cronbach's alpha = 0.85; person reliability = 0.84; item reliability = 0.97), and item difficulty spanned a broad range, enabling discrimination across students' ability levels. The novelty of this work lies in producing a Rasch-validated, constructed-response instrument explicitly aligned with the revised PISA 2025 science competency structure, including the new competency on evaluating and using scientific information in digital contexts. In conclusion, the instrument is suitable for limited classroom and research use. It provides a practical contribution to physics education by strengthening evidence-based assessment of scientific reasoning and decision-making in physics-related and environmental contexts.*

Keywords: *assessment instrument; science literacy; scientific reasoning; PISA 2025; Rasch model*

I. INTRODUCTION

Science literacy is a core competence for 21st-century learners because it enables individuals to explain natural phenomena, apply scientific knowledge to problem-solving, and make evidence-based decisions in personal and societal contexts (OECD, 2023). In contemporary life, scientific information is encountered in everyday situations, such as interpreting health claims, responding to environmental issues, and evaluating the risks and benefits of technology. These demands mean that learners should move beyond memorizing facts and toward using scientific concepts and methods to reason, interpret evidence, and justify decisions. Within PISA, science literacy is defined as the capacity to engage with science-related issues and ideas as a reflective citizen, highlighting the importance of learning and assessment that capture explanation, inquiry, and evidence-based reasoning (OECD, 2023).

International monitoring of science literacy is conducted through the Programme for International Student Assessment (PISA), which assesses the reading, mathematics, and science competencies of 15-year-old students at regular intervals. Indonesia has participated in PISA since 2000 (Yusmar & Fadilah, 2023); however, national performance has remained below the international benchmark. Indonesian students scored an average of 383 in science on the PISA 2022 test, which is lower than the OECD average of 485. This put Indonesia in 62nd place out of 81 countries that took the test (OECD, 2023). These results suggest difficulties with tasks that require interpreting data, evaluating evidence, and applying concepts to authentic contexts, indicating the need to strengthen both instruction and assessment in schools (OECD, 2023).

Evidence from diverse regions similarly indicates that low science literacy is not limited to a particular context. At the junior high school level, students in Bengkulu Province demonstrated very low science literacy achievement (Putri, 2021), while research in Langsa City, Aceh, reported weak science literacy skills, especially in creative and higher-order thinking (Oktaviani et al., 2023). This persistent pattern has been linked to several factors, including teachers' limited capacity to design science literacy-oriented assessments (Kamil et al., 2021; Oktaviani et al., 2023), curriculum demands that constrain instructional time (Fuadi et al., 2020), students' limited conceptual understanding and difficulties connecting concepts with real-world phenomena (Sutrisna, 2021; Permanasari, 2016), and inadequate learning facilities that restrict inquiry-based and experimental learning activities (Hidayah et al., 2019; Yusmar & Fadilah, 2023).

In classroom practice, these constraints can shape assessment routines that emphasize recall and procedural completion rather than the reasoning and decision-making dimensions of science literacy. When assessment tasks rarely require students to interpret evidence, evaluate information sources, or construct explanations and arguments, students receive limited opportunities to

practice the competencies targeted by PISA. Therefore, assessment instruments should be designed to align with competency indicators and supported by explicit scoring rubrics to ensure consistent scoring and diagnostically useful feedback (Kamil et al., 2021; Oktaviani et al., 2023).

To respond to this challenge, previous studies have proposed multiple interventions, such as implementing innovative learning models (Afriana et al., 2016; Azis et al., 2018; Saputra et al., 2022; Sari et al., 2017), developing science literacy-based teaching materials (Rostikawati & Permanasari, 2016; Rusyati et al., 2019), and strengthening assessment practices to measure students' competencies and inform instructional improvement. In particular, well-designed assessment instruments are crucial for diagnosing specific competency gaps and providing empirical feedback for learning design. Nevertheless, existing science literacy assessments in Indonesia still predominantly focus on junior high school students and general science contexts (Novitasari & Handhika, 2018). In contrast, senior high school students are expected to demonstrate more advanced competencies, such as scientific reasoning, investigation design, and critical interpretation of data, which are central to international assessments like PISA.

Development of science literacy assessment instruments has been conducted across varied science topics, including biological systems (Helendra & Sari, 2021), biodiversity (Alti et al., 2021), and environmental pollution (Martinah et al., 2022). Other studies have emphasized different orientations, such as AKM-based instruments (Saidah & Malichatin, 2023), contextual-based assessments (Martinah et al., 2022), and green chemistry-based instruments for senior high school students (Wirama, 2022). Although several studies have applied the Rasch Model and reported satisfactory validity and reliability, many instruments remain limited to specific subject domains and are aligned with earlier PISA frameworks. Consequently, they may not fully represent the integrated competencies and expanded information-evaluation demands introduced in the PISA 2025 framework.

The PISA 2025 framework introduces substantial revisions to science competency indicators. Two previously separate competencies, evaluating and designing scientific investigations and interpreting scientific data and evidence, are integrated into a single competency: constructing and evaluating scientific investigation designs and critically interpreting scientific data and evidence. In addition, a new competency is introduced in response to the increasing dominance of digital and internet-based information sources, namely researching, evaluating, and using scientific information for decision-making and action (OECD, 2023). These revisions underscore the need for science literacy assessment instruments that are explicitly aligned with the PISA 2025 framework and empirically validated using robust measurement models. Accordingly, this study aims to examine the validity and reliability of a

science literacy assessment instrument for senior high school students in the context of PISA 2025 using the Rasch Model.

II. METHODS

A research and development (R&D) approach was employed using the 4-D model: define, design, develop, and disseminate (Thiagarajan et al., 1974). This approach aimed to create a science literacy assessment instrument for measuring science competencies in the context of PISA 2025. The development process is summarized in Figure 1.

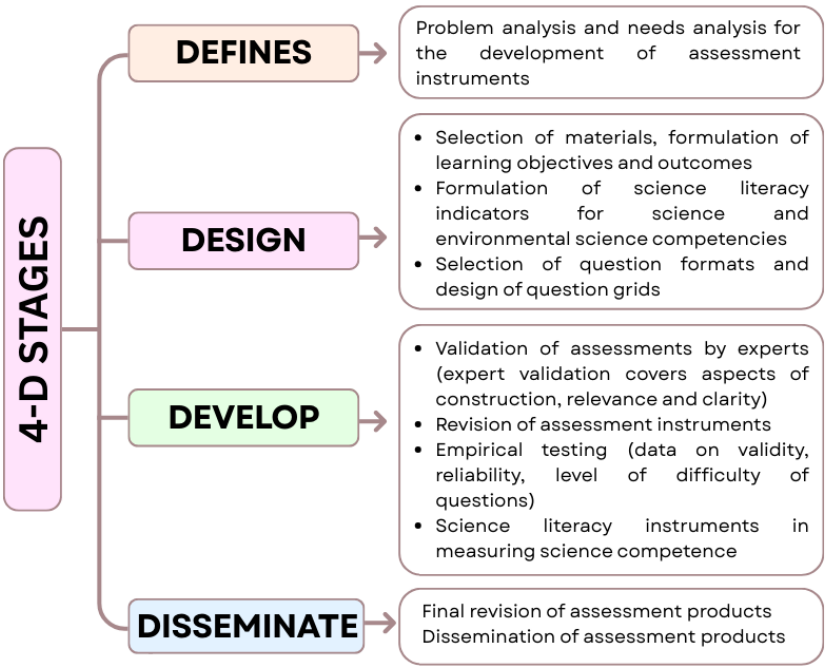

Figure 1. Research stages using the 4-D model

1. Define stage

During the Define stage, a needs analysis was conducted to identify key issues in evaluating science literacy in Langsa City. These issues included students' generally low achievement in science literacy, limited teacher training in science literacy, and teachers' difficulties in developing assessments aligned with science literacy, as reported in prior studies (Oktaviani et al., 2023). The findings from this stage informed the scope of competencies and the assessment format to be developed.

2. Design stage

In the design stage, curriculum and content analyses were carried out to select relevant materials, objectives, and learning outcomes aligned with the PISA 2025 science framework. Based on this alignment, indicators of science competence were formulated for three

competencies: (1) explaining phenomena scientifically; (2) constructing and evaluating scientific investigation designs and critically interpreting data and evidence; and (3) researching, evaluating, and using scientific information for decision-making and action. An essay format was selected to capture students' reasoning and evidence-based explanations consistent with PISA-style tasks. A test blueprint (item grid) and item scripts were then developed to ensure coverage of the defined indicators. Student responses were scored using a tiered analytical rubric with four performance levels (1–4) tailored to each item's characteristics (Table 1).

Table 1. Integrated essay scoring rubric based on PISA 2025 indicators

Score	Integrated descriptor
4	The response demonstrates accurate and comprehensive scientific understanding, clearly explains phenomena, critically evaluates scientific investigations and interprets data and evidence, and uses scientific information effectively to support well-justified decisions or actions in real-world contexts.
3	The response shows correct scientific explanations and reasonable interpretation of data and evidence, but evaluation of investigations or justification of decisions is limited or less comprehensive.
2	The response reflects partial understanding of scientific concepts, with superficial data interpretation and weak use of scientific information for decision-making.
1	The response is vague, inaccurate, or contains misconceptions, with minimal or no evidence of data interpretation, investigation evaluation, or evidence-based decision-making.

The rubric evaluated the accuracy of scientific concepts, depth of reasoning, quality of data interpretation, and the relevance and justification of proposed solutions or arguments. Higher scores indicated more comprehensive, scientifically accurate, and evidence-based reasoning, whereas lower scores reflected vague, incorrect, or misconception-based responses. To reduce rater subjectivity, the rubric included key indicators (keywords) and representative response exemplars for each performance level, thereby supporting transparent, replicable scoring aligned with the PISA 2025 framework.

3. Development stage

During the development stage, the initial instrument (ten PISA-oriented essay items) and the scoring rubric were refined and subjected to expert validation. A content validation questionnaire was prepared to evaluate three aspects: construction, relevance, and clarity. Using a 1–5 Likert scale. Three expert lecturers served as raters, and item-level content validity was quantified using Aiken's V:

$$V = \frac{\Sigma s}{n\,(c-1)}; \qquad s = r - lo \quad (1)$$

Explanation:

V = item validity index

s = r-lo

$\Sigma s = s1 + s2 + etc.$

n = number of raters

c = highest validity score

lo = lowest validity score

r = score given by a rater

Next, the overall score obtained will be interpreted in Table 1 to arrive at a final conclusion on whether the product assessed by the validator is valid or not. The interpretation of validity data against the validator's assessment is as follows (Aiken, 1985).

Table 2. Data interpretation of validity levels

No	Value	Value category
1	$0.00 < V \leq 0.20$	Very low
2	$0.20 < V \leq 0.40$	Low
3	$0.40 < V \leq 0.60$	Moderate
4	$0.60 < V \leq 0.80$	High
5	$0.80 < V \leq 1.00$	Very high

Following expert validation and revision, an empirical trial was conducted to assess the instrument's psychometric properties. The trial involved 50 randomly selected Year 11 senior high school students in Langsa City. Participants were selected by random sampling from one class in each of two senior high schools, ensuring that the sample represented a limited field trial rather than a large-scale implementation. The selected classes included students with heterogeneous academic abilities, reflecting typical classroom conditions and avoiding the bias associated with ability-based grouping. This sampling approach was considered appropriate for the initial empirical testing phase of a development study, where the primary objective is to evaluate item functioning and measurement quality rather than to generalize findings to a broader population. After data collection, Rasch Model analysis was employed to assess item validity, reliability, item difficulty, and discriminating power.

After expert validation and revision, a limited empirical trial was conducted to assess the instrument's psychometric properties. The trial involved 50 randomly selected Year 11 senior high school students in Langsa City. Participants were selected through random sampling from one class in each of two senior high schools, representing a limited field trial rather than large-scale implementation. The selected classes included students with heterogeneous academic abilities, reflecting typical classroom conditions. Students completed the ten essay items, and responses were scored using the analytic rubric (Table 1). Rasch Model analysis was then applied to evaluate item functioning and measurement quality, including item fit (Outfit MNSQ, Outfit ZSTD, and

Point Measure Correlation), item difficulty, and reliability indices for persons and items (Sumintono & Widhiarso, 2015).

III. RESULTS

The final instrument comprised 10 constructed-response (essay) items integrating science and environmental contexts aligned with the PISA 2025 framework. Table 3 presents the item guideline matrix used to ensure coverage across the targeted competencies.

Table 3. Science literacy skills question guidelines

No	Science competency aspects	Environmental science competency aspects	Item number
1	Explain phenomena scientifically	Explain the impact of human interactions with Earth's systems	1 and 2
2	Propose appropriate experimental designs	Search for and evaluate evidence from diverse knowledge systems and sources	3
3	Interpret data presented in different representations, draw appropriate conclusions from data, and evaluate their relative merits	Evaluate and design potential solutions to social, environmental and ecological issues using creative and systems thinking, taking into account implications for current and future generations	4, 5 and 6
4	Justify decisions using scientific arguments, either individual or communal, that contribute to solving contemporary issues or sustainable development	Respect diverse perspectives on issues and look for solutions to regenerate impacted communities and ecosystems	7, 8, 9 and 10
Total			**10**

Across the set, the contexts included sustainable energy supply, ecological contamination, climate change and global warming, transportation safety, and human–environment interactions affecting ecosystem sustainability. Student responses were scored using an analytic rubric with performance levels ranging from 0 to 4.

Table 4. Results of content validation for science literacy questions

Question number	Aspects			Interpretation
	Construction	Relevance	Clarity	
	V Value	V Value	V Value	
1	0.75	0.75	0.88	Valid
2	0.88	1.00	0.88	Valid
3	0.88	1.00	0.88	Valid
4	0.88	0.75	0.88	Valid
5	1.00	0.88	1.00	Valid
6	0.63	0.88	0.75	Valid

7	0.75	1.00	1.00	Valid
8	1.00	1.00	0.88	Valid
9	0.75	0.88	0.88	Valid
10	0.88	0.75	1.00	Valid
Mean	**0.84**	**0.89**	**0.90**	**Valid**

Content validation conducted by three expert lecturers resulted in high Aiken's V indices, indicating strong agreement on the evaluated aspects. The indices were 0.84 for construction, 0.89 for relevance, and 0.90 for clarity, all of which are considered indicative of excellent content validity. These indices are presented in Table 4 and demonstrate the robustness and reliability of the content as assessed by seasoned professionals. The consistency across the different criteria underscores the meticulous quality check, ensuring the material is well-constructed, relevant to the subject matter, and clearly articulated, thereby confirming its suitability for the intended purpose.

1. Item fit analysis

The evaluation of item fit was conducted utilizing Rasch Model criteria: Outfit MNSQ values ranging from 0.5 to 1.5, Outfit ZSTD scores between –2.0 and +2.0, and Point Measure Correlation coefficients spanning from 0.40 to 0.85 (Sumintono & Widhiarso, 2015). Items were retained if the MNSQ and Point Measure Correlation values satisfied the specified criteria, notwithstanding instances where ZSTD fell outside these ranges.

Table 5. Item validation results based on level of suitability (item fit)

Item	Outfit				Pt measure corr	Description	Decision
	MNSQ	Description	ZSTD	Description			
1	0.82	Accepted	-0.84	Accepted	0.55	Accepted	Valid
2	0.75	Accepted	-0.27	Accepted	0.41	Accepted	Valid
3	0.93	Accepted	-0.31	Accepted	0.63	Accepted	Valid
4	1.11	Accepted	0.59	Accepted	0.72	Accepted	Valid
5	0.93	Accepted	-0.31	Accepted	0.69	Accepted	Valid
6	0.66	Accepted	-1.92	Accepted	0.77	Accepted	Valid
7	1.29	Accepted	0.89	Accepted	0.55	Accepted	Valid
8	1.81	Rejected	2.53	Rejected	0.55	Accepted	Invalid
9	0.47	Rejected	-3.27	Rejected	0.79	Accepted	Invalid
10	0.72	Accepted	-1.47	Accepted	0.78	Accepted	Valid

Based on the item fit statistics shown in Table 5, eight out of ten items satisfied the Rasch fit criteria, indicating good model fit. However, Items 8 and 9 did not meet the Outfit MNSQ and ZSTD criteria, suggesting some misfit for these items. Despite this, their Point Measure Correlation values remained within the acceptable range, indicating a reasonable level of

correlation with the overall construct. These results highlight the overall adequacy of most items, with a few deviations requiring further review.

2. Instrument reliability

Reliability was examined using Cronbach's alpha as well as person and item reliability indices from the Rasch analysis.

SUMMARY OF 50 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	24.5	10.0	.88	.45	.97	-.25	.95	-.11
SEM	.9	.0	.18	.01	.12	.24	.12	.21
P.SD	6.3	.0	1.26	.06	.84	1.67	.85	1.48
S.SD	6.4	.0	1.27	.06	.85	1.69	.86	1.50
MAX.	38.0	10.0	4.29	.76	3.92	4.60	3.95	4.55
MIN.	9.0	10.0	-1.99	.40	.10	-3.37	.10	-2.21
REAL RMSE	.51	TRUE SD	1.15	SEPARATION	2.27	Person RELIABILITY	.84	
MODEL RMSE	.45	TRUE SD	1.17	SEPARATION	2.58	Person RELIABILITY	.87	
S.E. OF Person MEAN = .18								

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .85 SEM = 2.48
STANDARDIZED (50 ITEM) RELIABILITY = .97

SUMMARY OF 10 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	122.6	50.0	.00	.21	1.12	.14	.95	-.44
SEM	12.7	.0	.52	.02	.19	.77	.12	.51
P.SD	38.1	.0	1.56	.05	.56	2.30	.36	1.52
S.SD	40.1	.0	1.65	.05	.59	2.43	.38	1.60
MAX.	189.0	50.0	1.59	.33	2.29	4.73	1.81	2.53
MIN.	80.0	50.0	-3.15	.18	.49	-3.30	.47	-3.27
REAL RMSE	.25	TRUE SD	1.54	SEPARATION	6.24	Item RELIABILITY	.97	
MODEL RMSE	.21	TRUE SD	1.55	SEPARATION	7.31	Item RELIABILITY	.98	
S.E. OF Item MEAN = .52								

Figure 2. Person and Item Reliability

The reliability results for 10 items given to 50 students are quite positive! The Cronbach's alpha is 0.85, indicating good internal consistency. The person reliability is 0.84, showing solid consistency across individuals, and the item reliability stands at an impressive 0.97, reflecting very reliable items (see Figure 2).

3. Item difficulty level

Item difficulty was determined from the Rasch item measure (logit) values. The standard deviation (SD) of item measures in this test was 1.56; the difficulty categories used are summarized in Table 6.

Table 6. Categories of questions based on their level of difficulty

Logit value	Category
Greater than +1.56 SD	Very difficult
0.0 logit +1.56 SD	Difficult
0.0 logit -1.56 SD	Medium
Less than -1.56 SD	Easy

According to the established categories, Item 10 was identified as the most challenging; Items 9, 4, 6, 3, and 5 were classified as difficult; Items 1 and 8 as moderate; and Items 7 and 2 as easy (See Table 7)

Table 7. Item difficulty level

Entry number	Total score	Total count	JMLE measure	Interpretation
10	80	50	1.59	Very difficult
9	81	50	1.56	Difficult
4	90	50	1.27	Difficult
6	101	50	.92	Difficult
3	105	50	.78	Difficult
5	106	50	.75	Difficult
1	138	50	-.37	Medium
8	162	50	-1.36	Medium
7	174	50	-1.99	Easy
2	189	50	-3.15	Very easy

4. Person measure (students' ability level in answering items)

Student ability (person measure) was estimated in logit units. The SD of the person measures in this test was 1.26, and the grouping criteria are shown in Table 8.

Table 8. Criteria for grouping student ability

Student ability logit value	Category
Greater than +1.26	High
Less than +1.26	Medium
Less than -0.88	Low

As shown in Table 9, students 09P, 12P, and 04P had the highest logit estimates, while students 32L, 37L, and 49P had the lowest. Overall, 18 students were classified as high (36%), 29 as moderate (58%), and 3 as low (6%) in science literacy. These results indicate that most participants were in the moderate-to-high ability categories, with a small proportion at the low-ability level.

Table 9. Students' ability level in answering items

Person	JMLE measure range	Category
09P, 12P, 04P, 07P, 17L, 20L, 03P, 05P, 16L, 35L, 01P, 08L, 11P, 18L, 22P, 14P, 24P, 25P	1.44 – 4.29	High
02P, 23L, 48P, 19P, 33L, 26P, 34L, 50P, 39P, 43L, 40P, 36L, 42L, 44P, 45P, 13P, 15L, 27P, 29P, 47L, 10P,	-.64 – 1.23	Medium

28P, 30P, 46P, 21L, 41P, 31P, 06P, 38P		
32L, 37L, 49P	(-1.99) – (-1.15)	Low

5. Wright map (student ability and probability in answering items)

The Wright Map (Person–Item Map) in Figure 3 displays the distribution of person ability and item difficulty on the same logit scale, enabling direct comparison of respondent proficiency and item challenge.

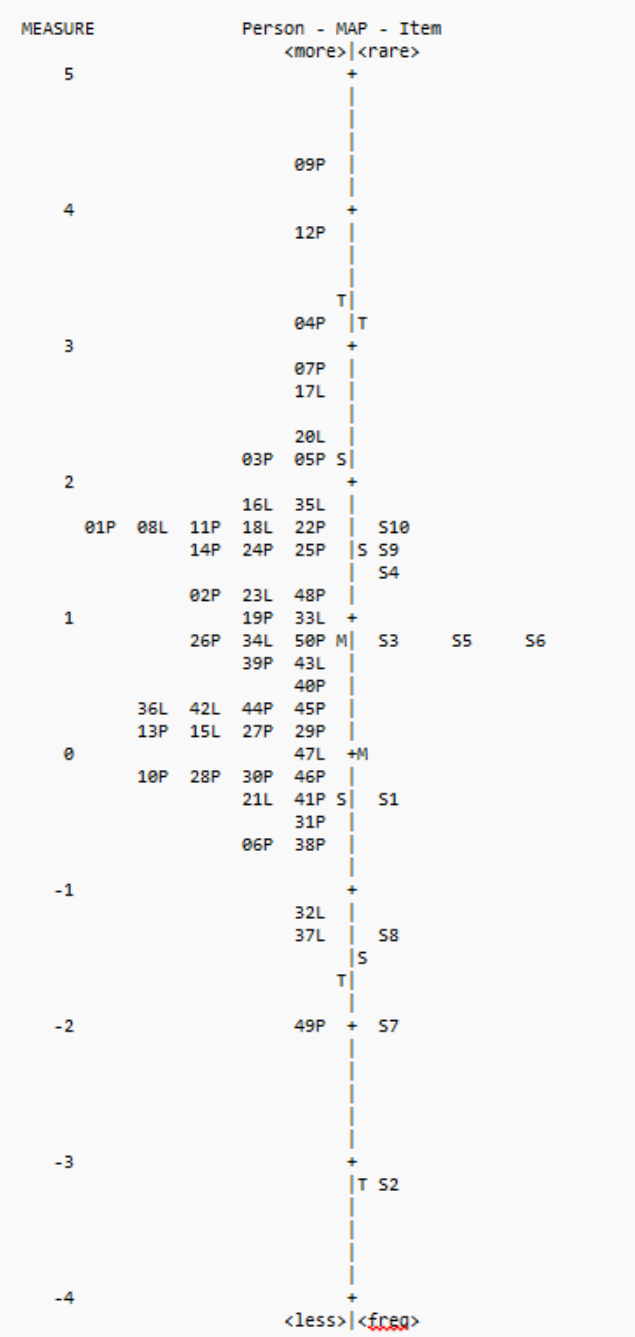

Figure 3. Variable (wright) map

In this sample, person ability estimates ranged from approximately -2 to +5 logits, with most students clustered between 0 and +2 logits. Item difficulty estimates ranged from approximately -3 to +3 logits, with the most difficult item located near +2 logits and the easiest near -3 logits. The person–item distribution shows that the instrument includes items spanning multiple difficulty levels and can differentiate students across a broad range of science literacy proficiency.

IV. DISCUSSION

The current study presents a comprehensive body of validity evidence for a science literacy assessment aligned with PISA 2025. This process commenced with expert-based content validation and proceeded through empirical evaluation of item functioning and reliability via Rasch modeling. Overall, the results demonstrate that the instrument is adequately valid and reliable for assessing the science literacy of senior high school students within the specified framework. Additionally, the findings identify particular domains for improvement to enhance measurement accuracy and sensitivity, especially for higher-ability students. The expert validation results demonstrate strong content validity across the three evaluated aspects. The average Aiken's V value for construction (0.84) indicates that the form, structure, and arrangement of items conform to established principles of good item writing, which is essential because construction quality influences item readability and reduces construct-irrelevant difficulty (Arikunto, 2013). In practice, well-constructed items support students in allocating cognitive resources to scientific reasoning rather than to deciphering item format or confusing prompts.

Regarding relevance, an average Aiken's V value of 0.89 confirms that the items align well with the intended science literacy indicators. This alignment is crucial because relevance ensures that the instrument measures the targeted competency domain rather than related or unintended constructs. In other words, strong relevance supports the representativeness of the science literacy construct as operationalized in the instrument, which is essential for making valid interpretations of student performance. The clarity aspect yielded the highest average Aiken's V value (0.90), indicating that item language and statements were judged understandable for students. Linguistic clarity is particularly important in literacy-oriented assessments because ambiguous wording can lead to multiple interpretations and introduce bias unrelated to the targeted skills (Sugiyono, 2017). In the context of PISA-type tasks often embedded in real-world scenarios, clarity helps ensure that performance differences reflect variation in science literacy rather than reading ambiguity or misinterpretation of contextual cues.

The magnitude of these Aiken's V indices is also consistent with validation practices reported in the literature. Several studies describe a critical threshold for item acceptance, such as 0.77 at the 95% confidence interval (García-Ceberino et al., 2020), and other empirical developments frequently report Aiken's V values above 0.80 for assessment instruments across educational contexts (Hidayah & Muhtarom, 2023). Taken together, the present results provide strong expert agreement that the instrument content is appropriate, representative, and clearly expressed. Despite these strong content validity outcomes, content validation alone does not guarantee that items function effectively when administered to students. Aiken's V evaluates alignment and suitability based on expert judgement. Still, it does not directly test whether students respond to items in a manner consistent with the intended measurement model, nor does it reveal how the items behave across the ability continuum. Therefore, empirical validation through pilot testing remains essential to examine item functioning, response consistency, and reliability in real testing conditions. In this study, the instrument was administered to 50 Year 11 students from two senior high schools in Langsa City, and responses were analyzed using Winstep software with the Rasch Model.

The item fit analysis showed that 8 of 10 items met the Rasch model fit criteria, suggesting that most items functioned consistently in measuring students' science literacy. Items that meet acceptable thresholds for Outfit MNSQ, ZSTD, and Point Measure Correlation typically indicate that response patterns align with model expectations and that item difficulty is meaningfully related to student ability. This supports construct validity from an internal-structure perspective and indicates that the instrument can be interpreted using the Rasch-calibrated scale. Similar work also reports that Rasch-based validation is effective for strengthening the psychometric quality of science literacy assessments, particularly when targeting higher-order competencies (Darman et al., 2024).

Items 8 and 9 exhibited misfits based on Outfit MNSQ and ZSTD, although their Point Measure Correlation coefficients remained within acceptable limits. This pattern implies that the items were still positively associated with the underlying construct but produced more variable response behavior than expected. In PISA-oriented assessments, such variability can occur when tasks require multi-step reasoning, such as interpreting evidence, evaluating an investigation design, or making decisions grounded in scientific information, because students may employ diverse solution strategies or be differentially affected by contextual complexity. Importantly, prior research cautions that misfitting items are not always "bad" items; they may reflect higher cognitive demand or sensitivity to student heterogeneity and are often best treated as candidates for targeted refinement rather than immediate removal (Adhari et al., 2025). In practical terms, refinement may include tightening the task prompt to reduce ambiguity, ensuring that contextual

information is necessary (not distracting), and strengthening rubric descriptors so that scoring captures intended reasoning patterns more consistently.

From an educational measurement perspective, the predominance of fitting items indicates that the developed instrument is generally suitable for assessing science literacy aligned with the PISA 2025 framework. Moreover, identifying misfitting items provides diagnostically valuable information to improve interpretability and strengthen the instrument's coherence. This approach aligns with recommendations that emphasize Rasch analysis not merely as a statistical test but as a tool to guide evidence-based improvement of PISA-oriented science assessments (OECD, 2023).

The reliability results further support the instrument's quality. Cronbach's alpha (0.85) indicates strong overall consistency in the interaction between persons and items, suggesting that the instrument yields stable measurement across students. Person reliability (0.84) indicates that the assessment can differentiate students' abilities with good consistency. In contrast, item reliability (0.97) suggests that the relative ordering of item difficulties is highly stable within the tested sample. These values exceed commonly used acceptability thresholds (e.g., >0.70) and are consistent with prior Rasch-based development studies reporting high person and item reliability for science literacy instruments (Adhari et al., 2025). They are also aligned with evidence that alpha values above 0.80 indicate reliable instruments in educational measurement contexts (Bashooir & Supahar, 2018). High reliability implies that students' response patterns are sufficiently consistent for the instrument to measure the targeted competencies effectively and to support meaningful interpretation of scores (Sumintono & Widhiarso, 2015).

The distribution of item difficulty from very easy to very difficult indicates that the instrument covers a broad range of challenge levels. This is a desirable property for PISA-oriented measurement, because PISA tasks are intentionally designed to span varying degrees of cognitive complexity and to assess learners from lower to higher proficiency levels. In this context, the more difficult items likely reflect advanced competencies central to PISA 2025, such as critically interpreting data, evaluating scientific investigations, and using evidence for decision-making (OECD, 2023). Conversely, easier and moderately difficult items remain essential for assessing foundational competencies (e.g., explaining scientific phenomena) and for obtaining informative measurement among students with lower levels of science literacy (OECD, 2019; Stacey, 2021).

This spread also supports the principle of item targeting, ensuring that items are appropriately aligned with respondents' ability distributions. Well-targeted instruments improve measurement precision and strengthen the interpretability of proficiency levels (Boone et al., 2014; Linacre, 2011). Additionally, Rasch-calibrated difficulty information is particularly valuable for refining PISA-like assessments because it helps identify gaps in item coverage and

supports systematic expansion of item pools to better capture complex, real-world literacy demands (Serdianus & Saputra, 2023). Therefore, the difficulty results suggest that the instrument is both psychometrically robust and pedagogically aligned with international science literacy assessment practices. Person measures indicate that most students in the sample demonstrate moderate to high science literacy skills, with a smaller proportion at low ability. Specifically, 18 students were classified as high (36%), 29 as moderate (58%), and 3 as low (6%). Students 09P, 12P, and 04P had the highest logit values, while students 32L, 37L, and 49P had the lowest. This distribution suggests that the instrument can differentiate students across levels and that, within this sample, many students can engage with the scientific contexts and demands represented by the items.

These findings are consistent with evidence from PISA–aligned studies showing that instruments grounded in PISA competencies can effectively separate students across ability bands and often exhibit a concentration at moderate-to-high levels, depending on sample characteristics (Zhang et al., 2023). The use of Rasch scaling also strengthens interpretability by placing students and items on a common logit metric, which supports more meaningful comparisons of ability relative to item difficulty and can inform instruction by identifying which difficulty regions best match the cohort.

Several limitations should be considered when interpreting these results. First, the pilot involved 50 students from two schools within one city, which constrains generalizability to broader populations. Second, the trial represents a limited field test; further studies with larger and more diverse samples are required to confirm the stability of item parameters and person estimates. Third, because the instrument uses constructed-response items, scoring consistency is critical; therefore, future work should include additional checks to ensure scoring robustness, and that rubric application remains consistent across contexts.

Despite these limitations, the results provide strong initial support for the instrument as a PISA 2025–oriented measure of science literacy. The high content validity and reliability, together with predominantly satisfactory Rasch fit, indicate that the instrument is suitable for diagnostic and evaluative purposes in similar educational settings (OECD, 2023; Tabacholly et al., 2025). For refinement, Items 8 and 9 should be prioritized for revision, focusing on prompt clarity, contextual load, and rubric specificity so that response patterns better conform to model expectations while preserving coverage of high-demand competencies (Adhari et al., 2025). Additionally, the observed targeting profile supports expanding the pool of high-difficulty items to characterize advanced proficiency better and strengthen alignment with the upper-range demands of PISA 2025 (OECD, 2023). Subsequent validation with larger samples will help consolidate the instrument's psychometric evidence base and support broader implementation.

V. CONCLUSION AND SUGGESTION

This study developed and validated a PISA 2025–oriented science literacy assessment instrument comprising 10 essay items, supported by an analytic scoring rubric. Expert judgement confirmed strong content validity across construction (Aiken's V = 0.84), relevance (Aiken's V = 0.89), and clarity (Aiken's V = 0.90), indicating that the items are well-formed, aligned with the intended indicators, and clearly understood. Empirical testing using Rasch analysis indicated that 8 of the 10 items met the model-fit criteria. In comparison, two items (Items 8 and 9) exhibited misfit on selected indices but still demonstrated acceptable point-measure correlations. The instrument also demonstrated high reliability (Cronbach's alpha = 0.85; person reliability = 0.84; item reliability = 0.97) and a broad range of item difficulty levels, supporting its use to measure students' science literacy across varying proficiency levels in the context of PISA 2025 competencies.

Despite these promising results, this study was limited by a small-scale trial involving 50 Year 11 students from two senior high schools in one city, which restricts the generalizability of the findings. Future research should involve larger, more diverse samples to confirm the stability of item parameters, strengthen evidence of construct validity, and examine item functioning across groups. Further work is also recommended to refine misfitting items (Items 8 and 9), expand the bank with more high-difficulty items to target advanced learners better, and reinforce scoring robustness through additional rater calibration and consistency checks for constructed-response scoring. Nevertheless, this study contributes to the field of physics education by providing a rigorously validated, PISA 2025–aligned science literacy assessment package that can support teachers and researchers in diagnosing students' scientific reasoning, evidence interpretation, and decision-making skills within real-world physics-related and environmental contexts, thereby strengthening assessment practices consistent with international standards.

REFERENCES

Adhari, D., Santiani, S., & Fatmawati, S. (2025). Validity and reliability analysis of an environmentally integrated science literacy test using Rasch model. *EduFisika: Jurnal Pendidikan Fisika, 10*(1), 38–53. https://doi.org/10.59052/edufisika.v10i1.42136

Afriana, J., Permanasari, A., & Fitriani, A. (2016). Penerapan project-based learning terintegrasi STEM untuk meningkatkan literasi sains siswa ditinjau dari gender. *Jurnal Inovasi Pendidikan IPA, 2*(2), 202–212. https://doi.org/10.21831/jipi.v2i2.8561

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Alti, R. P., Lufri, L., Helendra, H., & Yogica, R. (2021). Pengembangan instrumen asesmen berbasis literasi sains tentang materi keanekaragaman hayati kelas X. *Journal for Lesson and Learning Studies, 4*(1), 53–58. https://doi.org/10.23887/jlls.v4i1.34270

Arikunto, S. (2013). *Prosedur penelitian: Suatu pendekatan praktik*. Rineka Cipta.

Azis, A. A., Lutfi, & Ismail. (2018). Pengaruh project-based learning terintegrasi STEM terhadap literasi sains, kreativitas, dan hasil belajar peserta didik. *Prosiding Seminar Nasional Biologi dan Pembelajarannya*, 189–194. https://ojs.unm.ac.id/semnasbio/article/view/6984

Bashooir, K., & Supahar, S. (2018). Validitas dan reliabilitas instrumen asesmen kinerja literasi sains pelajaran fisika berbasis STEM. *Jurnal Penelitian dan Evaluasi Pendidikan, 22*(2), 219–230. https://doi.org/10.21831/pep.v22i2.19590

Boone, W. J., Yale, M. S., & Staver, J. R. (2014). *Rasch analysis in the human sciences*. https://doi.org/10.1007/978-94-007-6857-4

Darman, D. R., Suhandi, A., Kaniawati, I., Samsudin, A., & Wibowo, F. C. (2024). Development and validation of scientific inquiry literacy instrument (SILI) using Rasch measurement model. *Education Sciences, 14*(3), 1-28. https://doi.org/10.3390/educsci14030322

Fuadi, H., Robbia, A. Z., Jamaluddin, J., & Jufri, A. W. (2020). Analisis faktor penyebab rendahnya kemampuan literasi sains peserta didik. *Jurnal Ilmiah Profesi Pendidikan, 5*(2), 108–116. https://www.neliti.com/publications/431701/analisis-faktor-penyebab-rendahnya-kemampuan-literasi-sains-peserta-didik

García-Ceberino, J. M., Antúnez, A., Ibáñez, S. J., & Feu, S. (2020). Design and validation of the instrument for the measurement of learning and performance in football. *International Journal of Environmental Research and Public Health, 17*(13), 1-22. https://doi.org/10.3390/ijerph17134629

Helendra, H., & Sari, D. R. (2021). Pengembangan instrumen asesmen berbasis literasi sains tentang materi sistem ekskresi dan sistem pernapasan. *Jurnal Ilmiah Pendidikan Profesi Guru, 4*(1), 17–25. https://doi.org/10.23887/jippg.v4i1.29860

Hidayah, N., & Muhtarom, M. (2023). Validity and reliability test of teaching materials using Aiken's V formula and SPSS 22. *Schola, 1*(2), 75–82. https://doi.org/10.26877/schola.v1i2.342

Hidayah, N., Rusilowati, A., & Masturi, M. (2019). Analisis profil kemampuan literasi sains siswa SMP/MTs di Kabupaten Pati. *Phenomenon: Jurnal Pendidikan MIPA, 9*(1), 36–47. https://doi.org/10.21580/phen.2019.9.1.3601

Kamil, F. F., Permanasari, A., & Riandi, R. (2021). Studi profil literasi sains siswa dan pembelajarannya di SMP Kota Banda Aceh. *Jurnal IPA & Pembelajaran IPA, 5*(4), 353–363. https://doi.org/10.24815/jipi.v5i4.23446

Linacre, J. M. (2011). *A user's guide to Winsteps®: Rasch-model computer programs*. Winsteps.com. www.winsteps.com/manuals.htm

Martinah, A. A., Mubarok, V., Miarsyah, M., & Ristanto, R. H. (2022). Pengembangan instrumen tes literasi sains berbasis kontekstual pada materi pencemaran lingkungan. *Bioedusiana: Jurnal Pendidikan Biologi, 6*(2), 192–218. https://doi.org/10.37058/bioed.v6i2.3251

Novitasari, L., & Handhika, J. (2018). Profil analisis kebutuhan pengembangan instrumen kognitif literasi sains untuk siswa SMA. *Seminar Nasional Quantum, 25*, 517–523. https://www.semanticscholar.org/paper/Profil-analisis-kebutuhan-pengembangan-instrumen-Novitasari-Handhika/b1e86830b70d9e27aa18f2fb2d6670b094e42721

OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2023). *PISA 2025 science framework* (Second draft). https://pisa-framework.oecd.org/science-2025/assets/docs/PISA_2025_Science_Framework.pdf

Oktaviani, C., Seprianto, S., & Putri, M. D. (2023). Creative thinking-oriented students' scientific literacy skills: Preliminary study. *Jurnal Penelitian Pendidikan IPA, 9*(10), 8245–8250. https://doi.org/10.29303/jppipa.v9i10.5520

Permanasari, A. (2016). STEM education: Inovasi dalam pembelajaran sains. *Prosiding Seminar Nasional Pendidikan Sains,* 23–34. https://media.neliti.com/media/publications/173124-ID-stem-education-inovasi-dalam-pembelajara.pdf

Putri, M. D. (2021). Identifikasi kemampuan literasi sains siswa di SMP Negeri 2 Pematang Tiga Bengkulu Tengah. *Gravitasi: Jurnal Pendidikan Fisika dan Sains, 4*(1), 9–17. https://doi.org/10.33059/gravitasi.jpfs.v4i01.3610

Rostikawati, D. A., & Permanasari, A. (2016). Rekonstruksi bahan ajar dengan konteks socio-scientific issues pada materi zat aditif makanan untuk meningkatkan literasi sains siswa. *Jurnal Inovasi Pendidikan IPA, 2*(2), 156–164. https://doi.org/10.21831/jipi.v2i2.8814

Rusyati, R., Permanasari, A., & Ardianto, D. (2019). Rekonstruksi bahan ajar berbasis STEM untuk meningkatkan literasi sains dan teknologi siswa pada konsep kemagnetan. *Journal of Science Education and Practice, 2*(2), 10–22. https://doi.org/10.33751/jsep.v2i2.1395

Saidah, E. N., & Malichatin, H. (2023). Pengembangan instrumen literasi sains berbasis Asesmen Kompetensi Minimum (AKM) untuk peserta didik kelas VII SMP/MTs. *NCOINS: National Conference of Islamic Natural Science*, *3,* 240–255. https://www.collegesidekick.com/study-docs/28017374

Saputra, I. G., P., E., Sukariasih, L., Nursalam, L. O., & Desa, S. S. (2022). The effect of scientific literacy approach with discovery learning model toward physics concepts understanding. *Jurnal Pendidikan Fisika, 10*(2), 144–153. https://doi.org/10.26618/jpf.v10i2.7769

Sari, D. N. A., Rusilowati, A., & Nuswowati, M. (2017). Pengaruh pembelajaran berbasis proyek terhadap kemampuan literasi sains siswa. *PSEJ (Pancasakti Science Education Journal), 2*(2), 114. https://scienceedujournal.org/index.php/PSEJ/article/view/85

Serdianus, S., & Saputra, T. (2023). Peran artificial intelligence ChatGPT dalam perencanaan pembelajaran di era Revolusi Industri 4.0. *Masokan: Ilmu Sosial dan Pendidikan, 3*(1), 1–18. https://doi.org/10.34307/misp.v3i1.100

Stacey, K. (2021). Assessing mathematical and scientific literacy: Reflections from PISA. *Educational Studies in Mathematics, 108*(1–2), 5–24. https://link.springer.com/book/10.1007/978-3-319-10121-7

Sugiyono. (2017). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Alfabeta.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Trim Komunikata Publishing House.

Sutrisna, N. (2021). Analisis kemampuan literasi sains peserta didik SMA di Kota Sungai Penuh. *Jurnal Inovasi Penelitian, 1*(12), 2683–2694. https://www.neliti.com/publications/469443/analisis-kemampuan-literasi-sains-peserta-didik-sma-di-kota-sungai-penuh

Tabacholly, F. A., Susongko, P., & Nafiati, D. A. (2025). Construct validation of science literacy instrument with Rasch modeling on students of Grade VIII junior high school. *Journal of*

English Language and Education, 10(1), 378–388. https://doi.org/10.31004/jele.v10i1.652

Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). *Instructional development for training teachers of exceptional children*. Leadership Training Institute/Special Education, University of Minnesota. https://www.semanticscholar.org/paper/InstructionalDevelopment-for-Training-Teachers-ofThiagarajan/44a718a0c8e219b37aabb4c36117dcd695c895d0

Wirama, T. G. P. (2022). Asesmen literasi sains tema kimia hijau pada siswa. *Indonesian Journal of Educational Development, 3*(1), 1–15. https://doi.org/10.5281/zenodo.6557263

Yusmar, F., & Fadilah, R. E. (2023). Analisis rendahnya literasi sains peserta didik Indonesia: Hasil PISA dan faktor penyebab. *LENSA (Lentera Sains): Jurnal Pendidikan IPA, 13*(1), 11–19. https://doi.org/10.24929/lensa.v13i1.283

Zhang, L., Liu, X., & 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*(21), 1-15. https://doi.org/10.1186/s43031-023-00093-2