AlphaMath
Journal of Mathematics Education, 12(1) May 2026
p-ISSN: 2477-409X, e-ISSN: 2549-9084
Available online: https://jurnalnasional.ump.ac.id/index.php/alphamath/

Mathematical Literacy Assessment: A Scalable Mobile Adaptive Blueprint for Mapping Proficiency Across PISA Domains

Ilham Falani*, Syamsir Sainuddin
Universitas Jambi, Indonesia
ilhamfalani@unja.ac.id

http://dx.doi.org/10.30595/alphamath.v12i1.30279

ABSTRACT

Indonesian students continue to struggle with mathematical literacy, as demonstrated across several cycles of PISA assessments. This study addresses this gap by applying a Rasch-based diagnostic approach to map patterns of item difficulty and examine how students across different grade levels engage with key indicators and content domains. A total of 271 students in Grades VII to IX from Indonesian public schools completed a 32-item multiple-choice test aligned with the PISA indicators of Formulating, Employing, Interpreting, and Reasoning. The instrument was specifically designed to capture different cognitive levels and situational contexts within the mathematical literacy framework. Rasch analysis was used to evaluate person ability, item difficulty, model fit, and measurement assumptions. The model showed strong empirical evidence, with a person reliability of 0.85, item reliability of 0.98, and infit/outfit MNSQ values within the acceptable range (0.5 to 1.5), and no significant Differential Item Functioning was found across grade levels. Formulating was the most difficult indicator, while Interpreting was the easiest. Among the content domains, Space and Shape posed the greatest challenge, whereas Quantity was the most accessible domain. Grade VIII students demonstrated the highest mean ability, producing a non-linear pattern across grade levels, likely due to a shift in Grade IX toward procedural exam-oriented instruction that narrows the focus on high-order modeling skills. Findings suggest that difficulties in modelling and spatial reasoning arise from deeper conceptual issues rather than grade progression alone. These results highlight the need for instructional practices that place greater emphasis on modelling processes and spatial reasoning.

Keywords: Diagnostic Assessment, Mathematical Literacy, PISA, Rasch Model.

Received : April 1, 2026
Revised : May 17, 2026
Accepted : May 26, 2026

Introduction

Mathematical literacy continues to receive considerable attention in Indonesia because many students still experience difficulties applying mathematical concepts in unfamiliar or real-world contexts. The most recent Programme for International Student Assessment (PISA) 2022 results indicate that Indonesian 15-year-old students achieved an average mathematics score of approximately 366 points, substantially below the Organization for Economic Co-operation and Development (OECD) average of 472 points, with only a small proportion of students reaching proficiency Levels 5 or 6 (OECD, 2023b). These upper levels require learners to analyse complex information, construct meaningful models, and evaluate the reasonableness of solutions. In daily classroom practice, teachers often observe that students can perform procedures they have memorised, but hesitate when asked to interpret data, identify relationships, or decide how a context can be translated into mathematical form (Aulia et al., 2024). This persistent gap between curriculum expectations and students' conceptual readiness has stimulated growing discussions regarding the improvement of mathematical literacy as a foundation for strengthening reasoning and problem-

solving in Indonesian schools (Ma'ruf et al., 2024; Ministry of Education, Culture, and Research, 2024).

The PISA 2022 framework defines mathematical literacy as the capacity to reason mathematically and to formulate, employ, and interpret mathematics across personal, occupational, societal, and scientific contexts (OECD, 2023a). Within this framework, content knowledge is organised into four categories: Quantity, Change and Relationships, Space and Shape, and Uncertainty and Data. Research shows that many students struggle to recognize mathematical structure when information is presented in textual, tabular, or visual forms. Previous studies report that students frequently rely on superficial cues when solving contextual problems (Risdiyanti et al., 2024), and have documented various errors that occurr not during computation but during the mathematization process (Gedik Altun & Morkoyunlu, 2023). These findings indicate that challenges may not be evenly distributed across PISA indicators or content domains, particularly in tasks involving modelling and spatial reasoning.

To better understand these conceptual barriers, several key studies have mapped specific failure points in students' literacy performance. The synthesis of literature presented in Table 1 highlights a consistent breakdown in the mathematical modelling cycle among students. This difficulty is often exacerbated by a reliance on superficial cues rather than deep semantic analysis of the problem, a trend documented by (Risdiyanti et al., 2024). Furthermore, the errors occurring during this "mathematization" phase are systematic rather than accidental, reflecting a gap in students' ability to transform contextual constraints into formal representations (Gedik Altun & Morkoyunlu, 2023). By aligning these empirical challenges with the (OECD, 2023a) framework, it becomes evident that diagnostic efforts must move beyond scoring and toward mapping specific indicator-based hurdles. These studies collectively support the need for the current research, which utilizes Rasch modeling to provide a more precise diagnostic map of where these conceptual breakdowns occur across different grade levels.

The growing interest in diagnostic assessment reflects the need to understand students' thinking in greater depth than can be captured by total test scores alone. Rasch modelling offers a promising approach because it aligns item difficulty and student ability on a common interval scale and provides evidence about how well items represent the underlying construct (Bond et al., 2021; Wind & Hua, 2022b). Studies using Rasch analysis in Indonesian mathematics education demonstrate its value for identifying patterns of difficulty and evaluating the quality of test items. (Oliva & Blanco, 2023), for instance, found that Rasch analysis helps reveal misfitting items and structural issues in an assessment, thereby clarifying where students'

p-ISSN 2477-409X, e-ISSN: 2549-9084

conceptual understanding does not align with the intended construct. Likewise, Rasch modelling has been used to track how students respond to tasks across different difficulty levels, showing identifying points at which they struggle to progress from basic to more advanced forms of reasoning (L'Boy & Nazim Khan, 2023). These contributions support the use of modern measurement models to examine how students engage with mathematical literacy tasks. The limited number of Rasch-based studies that specifically analyse PISA-aligned items in Indonesia created a clear starting point for the present research, which focuses on identifying the indicators and content areas that present the greatest conceptual challenges for lower-secondary students.

Rasch modelling has become a central tool in this area because it locates item difficulty and student ability on a common interval scale, enabling more precise evaluation of how well test items target learners' abilities (Bond et al., 2021; Wind & Hua, 2022b). Indonesian studies have used this model to examine the quality of mathematics assessments and the alignment between items and learners. Rasch analysis has been applied to algebraic thinking tests to evaluate item fit, estimate item difficulty, and map the precise correspondence between item demands and student performance (Rusyid et al., 2024). Similarly, Rasch modeling has been used to determine the difficulty levels and suitability of fraction test items, providing detailed information regarding which items are appropriate for students across various ability levels (Karlimah, 2022). International studies echo these findings, indicating that Rasch-based diagnostic analyses can reveal uneven development and problematic items that are not readily visible through descriptive statistics alone. Taken together, these studies underline the value of modern measurement models in analysing how learners interact with assessment tasks, including those aligned with PISA-style mathematical literacy.

Research on mathematical literacy in Indonesia consistently shows that many students experience difficulties when engaging with context-based tasks. Much of this difficulty stems from how students interpret problem situations and translate contextual information into mathematical form. One study reported that lower-secondary students showed uneven performance across different content domains of mathematical literacy, particularly when tasks required connecting real-world information with mathematical representations (Nurjanah et al., 2023). Evidence from younger learners points to a similar trend. These findings suggest that students' struggles are not confined to a single domain or grade level but may reflect deeper issues in how students reason about contextual information. Although this body of work offers valuable insights, empirical evidence remains limited regarding how these

difficulties vary across grade levels within lower-secondary education or how they relate to the cognitive processes defined in the PISA framework.

While previous research has provided useful descriptions of student errors, most studies focus primarily on surface-level errors and do not provide a clear understanding of which content domains pose the greatest overall challenges. Because mathematical literacy tasks encompass a wide range of concepts, from quantity and change to space and shape, a more structured understanding of domain-level difficulties is needed. In addition to content knowledge, mathematical literacy encompasses three cognitive processes: Formulating, Applying, and Interpreting, which shape how students understand a problem from start to finish. However, existing studies rarely locate student difficulties within these specific processes, making it difficult to discern whether the primary barriers lie in modeling situations, selecting and applying procedures, or interpreting results. While previous research has documented several types of difficulties, most studies have focused on isolated skills or single grade levels. What remains unclear is how these difficulties evolve as students progress through lower secondary education, whether they increase steadily, stagnate, or even fluctuate across grade levels. Understanding this development is crucial for identifying the points where instructional support is most needed. Possible research questions include:

RQ1: Which PISA process indicators present the greatest conceptual difficulties for Indonesian junior high school students?
RQ2: How do patterns of difficulties across PISA content domains reflect specific conceptual challenges in students' reasoning and modeling?
RQ3: How do Indonesian students' mathematical literacy difficulties vary across grade levels in junior high school?

Methods

Research Design

This study used a diagnostic measurement design supported by Rasch modelling to examine how students engaged with PISA-aligned mathematical literacy tasks. The purpose of this design was to identify patterns of conceptual difficulty across indicators, content domains, and grade levels, rather than to compare instructional treatments or to test causal relationships. Rasch modelling was chosen because it provides a coherent way to position item difficulty and student ability on a shared interval scale, which helps reveal subtle patterns in how students respond to different types of items (Ramalisa et al., 2023). Recent methodological work also supports the use of Rasch-based diagnostics when researchers aim to understand how learners interact with specific cognitive processes within large-scale assessment frameworks

Department of Mathematics Education, Universitas Muhammadiyah Purwokerto, Purwokerto, Indonesia
p-ISSN 2477-409X, e-ISSN: 2549-9084

(Aryadoust et al., 2020; Truong et al., 2024). This approach was well aligned with the objectives of the study, which focused on identifying where conceptual challenges were most prominent and whether these challenges varied across grade levels.

Participants

The study involved 271 students from Grades VII, VIII, and IX enrolled in three urban public junior high schools in Indonesia. The sample included 146 female students and 125 male students. There were 61 students in Grade VII, 89 in Grade VIII, and 121 in Grade IX. The schools were selected purposively based on three considerations. First, they shared a similar accreditation level and followed a comparable instructional structure, which ensured that students were exposed to standardised learning conditions. Second, all participating schools had stable enrolment and consistent participation in national assessments, making them suitable for a diagnostic study. Third, school administrators were willing to accommodate the data collection schedule during regular class hours.

Participant selection within each school followed an intact-class approach. Mathematics teachers identified classes whose schedules aligned with the available testing time, and all students present on the day of assessment were invited to participate. No financial incentives were provided to students. Instead, each school issued a certificate of participation that students could include as part of their academic portfolio. Because all participants were younger than 18 years of age, written institutional consent was obtained from the school principals, and teachers verbally confirmed that parents had been informed about the study. Ethical approval for this research was granted by the university's Educational Research Ethics Committee.

Research Instrument

The research instrument was a 32-item multiple-choice test aligned with the PISA 2022 mathematical literacy framework (OECD, 2023a). The items were operationalized to measure four process indicators (Formulating, Employing, Interpreting, and Reasoning) and four content domains (Quantity, Change and Relationships, Space and Shape, and Uncertainty and Data). Content validity of the instrument was established through expert judgment by three specialists in mathematics education, psychometrics, and instructional design. The first validator was a mathematics education expert with more than twelve years of experience in designing contextual and modelling-based tasks. The second validator was a measurement and assessment specialist with over ten years of experience applying Rasch modelling in educational research. The third validator was a curriculum and instructional design expert who had worked for more than fifteen years on the development of mathematics learning materials. They examined the clarity of item statements, the accuracy of the

mathematical content, and the alignment between each item and its intended PISA indicator. The resulting Aiken's V coefficients ranged from 0.83 to 0.94, exceeding the threshold for satisfactory content relevance. For items adapted from English sources, a rigorous back-translation procedure was implemented to preserve cognitive demand and linguistic equivalence in the Indonesian context.

Research Procedures

The study was conducted in four distinct stages. First, a preliminary pilot test (n=45) was performed to refine item wording and visual representations. The participants for this pilot stage were drawn from the same student population (Grades VII-IX) but from different schools than those in the main study, and this group was strictly excluded from the final sample. Second, the finalized instrument was administered during regular school hours under standardized conditions, with a 90-minute time allocation. Third, for items translated from English, a back-translation procedure was used to ensure that meaning and cognitive demand were preserved in the Indonesian version. Finally, raw responses were dichotomously coded (0 for incorrect, 1 for correct) and digitized for further psychometric evaluation.

Data Collection Techniques

Data were collected in situ at the participating schools over a scheduled two-week period. The assessment was administered by the research team in collaboration with mathematics teachers to maintain a naturalistic yet controlled classroom environment. To ensure the accuracy of the digitizing process, a double-entry data method was employed, where two independent researchers entered the raw scores into a digital database to minimize entry errors.

Data Analysis Techniques

Psychometric analysis was conducted using WINSTEPS version 5.1.0. The Rasch analysis followed a systematic modeling sequence to ensure measurement stability and diagnostic accuracy. Measurement Assumptions: Prior to estimation, unidimensionality was evaluated via Principal Component Analysis (PCA) of standardized residuals to ensure the main dimension accounted for the majority of the variance. Local independence was verified by inspecting item-residual correlations to confirm that responses to one item did not influence another. Fit Analysis: Infit and Outfit Mean Squares (MNSQ) values were evaluated using an acceptable range of 0.5 and 1.5. This ensures that the items function as expected and are productive for measurement (Bond et al., 2021). Reliability and Separation: The analysis assessed Person/Item Reliability and Separation indices, alongside KR-20, to determine the stability of the instrument and its capacity to distinguish among different levels of student ability. Invariance Testing: Differential Item Functioning (DIF) analysis was

p-ISSN 2477-409X, e-ISSN: 2549-9084

conducted across grade levels (VII, VIII, and IX). This was essential to verify parameter invariance, ensuring that item difficulty remained stable regardless of the students' schooling stage and that the test was fair for all subgroups (Bond et al., 2021; Wind & Hua, 2022a). Visual and Diagnostic Comparison: Wright Maps were generated to visualize the alignment between student ability (θ) and item difficulty (β). Additionally, mean logit values were calculated for each PISA indicator and content category to pinpoint specific areas of conceptual weakness.

Result and Discussions

Review of Rasch Measurement Assumptions

Before analysing the item and person estimates, the Rasch assumptions were checked to ensure that the measurement model behaved as expected. These assumptions relate to unidimensionality, local independence, and the stability of item functioning across groups. Verifying them is important because weak assumptions can distort how item difficulty or student ability is interpreted in diagnostic studies (Aryadoust et al., 2020).

Unidimensionality

Unidimensionality testing was conducted to ensure that the instrument measured a single underlying construct, namely, mathematical literacy. In the Rasch model, this assumption is fundamental as all items must function collectively to measure a specific variable. The researcher utilized a Principal Component Analysis (PCA) of residuals to verify this requirement. This analysis aims to identify whether any secondary dimensions are being measured alongside the primary construct. The results of the variance analysis are presented in Table 1.

Table 1. Standardized Residual Variance (Eigenvalue Units)

Variance Component	Eigenvalue	Empirical %	Modeled %
Total raw variance in observations	47.3	100.00	100.00
Raw variance explained by measures	15.3	32.40	31.00
Raw variance explained by persons	6.4	13.40	12.80
Raw variance explained by items	9	19.00	18.10
Raw unexplained variance (total)	32	67.60	69.00
Unexplained variance in the 1st contrast	2.6	5.50	8.20
Unexplained variance in the 2nd contrast	1.9	4.10	6.00
Unexplained variance in the 3rd contrast	1.8	3.80	5.60
Unexplained variance in the 4th contrast	1.6	3.30	4.90
Unexplained variance in the 5th contrast	1.5	3.10	4.60

Table 1 presents the results of the PCA of residuals used to verify the unidimensionality of the instrument. The PCA results show that the total raw variance explained by the measures was 32.40%. This value exceeds the minimum threshold of

20%, indicating that the instrument has a strong primary dimension. Furthermore, the unexplained variance in the first contrast was 5.50% with an eigenvalue of 2.6. This eigenvalue was significantly lower than the critical limit of 3.0. These findings demonstrate the absence of dominant secondary dimensions within the test. The noise or unexplained variance remains at a tolerable level, as it does not form a significant independent cluster of items. Consequently, the instrument meets the unidimensionality requirement, and the total score accurately reflects a single underlying trait of the students' mathematical literacy.

Local Independence

The researcher conducted a local independence test after confirming the unidimensionality of the instrument. This test aims to verify that a student's response to a particular item does not influence their response to other items. In the Rasch model, each item must provide unique and independent information regarding the student's proficiency. If two items are excessively correlated, the data is considered redundant, which may compromise measurement accuracy. The researcher used residual correlation values to assess the degree of dependence between these items. A summary of the largest residual correlations is presented in Table 2.

Table 2. Largest Standardized Residual Correlations

Correlation	Item 1 (Number & Code)	Item 2 (Number & Code)
0.27	01 – 1FCPQ	03 – 3ICCC
0.25	16 – 16R2SS	24 – 24R4OU
-0.25	24 – 24R4OU	31 – 31I6CC
-0.25	15 – 15I2SS	24 – 24R4OU
-0.25	05 – 5FBSS	13 – 13F2CC
-0.23	16 – 16R2SS	19 – 19I3OU
-0.23	14 – 14E2OU	20 – 20R3CC
-0.23	22 – 22E4PQ	24 – 24R4OU
-0.21	02 – 2ECPQ	22 – 22E4PQ
-0.21	12 – 12RAPQ	24 – 24R4OU

Residual correlations were inspected to verify that a student's response to one item did not depend on their response to another. Based on the data in Table 2, the highest standardized residual correlation was 0.27, identified between Item 01 and Item 03. This value remains below the commonly accepted critical threshold of 0.30. Further analysis shows that the remaining item correlations range from -0.25 to 0.25, with most values close to zero.

Department of Mathematics Education, Universitas Muhammadiyah Purwokerto, Purwokerto, Indonesia
p-ISSN 2477-409X, e-ISSN: 2549-9084

These findings suggest that each item measures a distinct aspect of the construct while remaining interrelated. Such patterns are typically observed in contextual mathematics assessments where tasks may share a common theme or stimulus but still function independently once student ability is accounted for (Lee & Yeo, 2022; Susanta et al., 2023). Consequently, the instrument satisfies the local independence requirement. This ensures that the resulting scores accurately reflect students' mathematical literacy proficiency without interference from item-to-item dependencies or redundancy.

Parameter Invariance

To examine whether items behaved similarly across subgroups, Differential Item Functioning (DIF) analysis was conducted for Grades VII, VIII, and IX. None of the items showed statistically significant DIF ($p > 0.05$). This indicates that differences in student performance were driven by ability rather than grade level, and that the relative ordering of item difficulty was stable across groups. Such stability is essential in diagnostic studies because it supports fairness and the comparability of results across populations (Aryadoust et al., 2020; Karlimah, 2022). Taken together, evidence from the dimensionality check, residual correlations, and DIF analysis suggests that the Rasch assumptions were adequately met for this dataset. The measurement structure was sufficiently coherent, which provides a sound basis for interpreting the person and item estimates in the next sections.

Instrument Validity and Reliability

Once the Rasch assumptions were met, the researcher evaluated the overall quality of the instrument and the interaction between students and items. This analysis aimed to verify the consistency of measurement results and ensure that the data aligned with the model's predictions. The researcher examined reliability values, separation indices, and fit statistics to determine the level of confidence in the test results. Overall, the summary statistics demonstrate a stable measurement pattern consistent with expectations in diagnostic studies of mathematical literacy. A summary of the fit statistics for both students and items is presented in Table 3.

Table 3 presents the summary of fit statistics for 271 students and 32 items. The student reliability is 0.85, while the item reliability reaches 0.98. Furthermore, the Cronbach's Alpha (KR-20) value of 0.88 indicates that the overall interaction between students and items demonstrates excellent internal consistency. Based on the data in Table 3, the separation index for students is 2.38, and for items it is 6.50. A separation index greater than 2.0 indicates that the instrument effectively differentiates between student proficiency groups and item difficulty levels. Additionally, the mean Outfit Mean Square (MNSQ) for both students and items is 1.15. This value falls within the ideal

range near 1.0, proving that the observed data aligns well with the Rasch model. This suggests that no extreme distorted response patterns were found, making the measurement results valid for mapping students' mathematical literacy proficiency.

Table 3. Summary of Fit Statistics

Statistics	Student (N=271)	Item (N=32)
Measures (logit)		
Mean	0.08	0.00
SE (standard error)	0.43	0.15
SD (standard deviation)	1.17	1.04
Outfit mean square		
Mean	1.15	1.15
SD	0.74	0.79
Separation	2.38	6.50
Reliability	0.85	0.98
Cronbach's Alpha (KR-20)	0.88	NA*

The mean person measure was close to zero, indicating that the overall item difficulty was well-matched to the students' ability range. A person separation index of 2.38, together with a reliability value of 0.85, suggests that the instrument was able to distinguish several ability strata within the sample, a level considered acceptable for classroom-based literacy assessments (Aryadoust et al., 2020; Boone, 2016). Item performance also appeared stable. The item separation index exceeded 6, and reliability was 0.98, which means that the ordering of item difficulty would likely remain consistent across similar samples. This high replicability is common in Rasch modelling when items draw on clearly defined cognitive indicators (Truong et al., 2024). The internal consistency of the test, reflected by a Cronbach's Alpha (KR-20) value of 0.88 (see Table 6, Student column), indicates that the items worked together cohesively to measure students' mathematical literacy. It should be noted that while Item Reliability (0.98) reflects the stability of the item difficulty hierarchy, the KR-20 represents the reliability of the person sample based on the internal consistency of responses across the instrument.

The researcher conducted an item fit analysis to evaluate whether each question functioned as expected under the Rasch model's predictions. This analysis is essential for identifying items that might confuse students or measure mathematical literacy inconsistently. The researcher utilized two primary indicators: the Outfit Mean Square (MNSQ) and the Point-Measure Correlation (Pt-Mea Corr). An item is considered to have an acceptable fit if the MNSQ value falls within the range of 0.5 to 1.5 and possesses a positive correlation value. The fit analysis results for the 32 items are presented in Table 4.

Department of Mathematics Education, Universitas Muhammadiyah Purwokerto, Purwokerto, Indonesia
p-ISSN 2477-409X, e-ISSN: 2549-9084

Table 4. Item Fit Analysis

Item Code	Measure	Infit MNSQ	Outfit MNSQ	Outfit ZSTD	Pt. Mea. Corr
01FCPQ	-0.59	0.793	0.708	-3.119	0.613
02ECPQ	-1.34	0.863	0.725	-1.849	0.516
03ICCC	-1.59	0.874	0.693	-1.799	0.493
04RCSS	-0.37	0.866	0.814	-2.109	0.568
05FBSS	0.27	1.106	1.101	1.211	0.408
06EBOU	-0.37	1.070	1.035	0.401	0.425
07IBOU	-0.91	0.904	0.823	-1.469	0.514
08RBPQ	-0.29	1.091	1.078	0.881	0.409
09FAOU	0.10	0.838	0.779	-2.889	0.602
10EACC	0.20	1.244	1.271	3.081	0.308
11IAPQ	-0.47	0.881	0.793	-2.259	0.560
12RAPQ	-0.94	0.915	0.765	-1.989	0.518
13F2CC	1.25	1.165	1.256	1.911	0.321
14E2OU	0.88	1.140	1.224	2.071	0.358
15I2SS	-0.63	0.877	0.767	-2.359	0.559
16R2SS	2.37	1.586	4.073	7.504	-0.292
17F3CC	2.65	1.322	3.607	5.804	-0.104
18E3CC	-0.35	0.847	0.764	-2.759	0.589
19I3OU	-0.79	0.817	0.716	-2.689	0.587
20R3CC	-1.20	0.858	0.709	-2.149	0.536
21F4OU	-0.06	1.013	0.955	-0.519	0.479
22E4PQ	-0.23	0.881	0.807	-2.319	0.568
23I4PQ	-0.35	0.855	0.784	-2.499	0.581
24R4OU	2.01	1.530	2.399	5.232	-0.087
25F5SS	0.99	1.172	1.404	3.341	0.314
26E5SS	-0.43	0.979	0.935	-0.659	0.487
27I5SS	-0.37	0.875	0.799	-2.289	0.567
28R5OU	-0.33	1.004	0.982	-0.169	0.470
29F6SS	-0.57	0.801	0.705	-3.189	0.611
30E6SS	1.92	1.315	1.982	4.172	0.101
31I6CC	-0.29	0.804	0.739	-3.159	0.617
32R6CC	-0.21	0.796	0.752	-3.079	0.621

Bolt and italic are the items not in the range of Outfit MNSQ and Pt-Measure Corr.

Table 4 presents the fit statistics for each item integrated into the application. Most items demonstrate Infit and Outfit MNSQ values within the recommended range. This indicates that the majority of the items have adequate quality for measuring student proficiency. However, a small number of items show misfit, particularly those with high difficulty levels, such as codes 16R2SS, 17F3CC, and 24R4OU.

These specific items exhibit Outfit MNSQ values exceeding the 1.5 threshold and

negative Point-Measure Correlation values. This condition suggests that these items do not align as closely with the primary construct as the other questions. Similar patterns are frequently observed in Rasch-based evaluations of PISA-style tasks, where highly contextual or multi-step modelling items tend to produce variable fit statistics (Hayat et al., 2020; Risdiyanti et al., 2024). Nevertheless, these deviations do not undermine the overall measurement structure. The researcher can refine these items by clarifying the context or simplifying the representational demands.

Taken together, the reliability, separation, and fit indices in Table 4 indicate that the 32-item instrument functions well for diagnostic purposes. The test demonstrates a good balance of item difficulty, strong internal coherence, and adequate discrimination across student abilities. These findings provide a solid basis for interpreting the subsequent analysis of item difficulty patterns and student performance across PISA domains.

Person-Item Distribution (Wright Map Analysis)

To understand how well the test targeted the range of student abilities, the Rasch estimates were examined through a Wright Map. A quick look at the distribution shows that most students were located near the middle of the ability continuum, while a smaller number were spread toward both extremes. The item difficulties covered a narrower range, with most items clustered between moderate and slightly challenging regions. This pattern is fairly typical for instruments designed to measure mathematical literacy, where items often concentrate around the central portion of the difficulty scale to capture variation in the majority of learners (Risdiyanti et al., 2024). One feature that stands out is the gap between the most difficult items and the ability levels of the higher-performing students. Items above +2 logits were rarely answered correctly, suggesting that they required a level of abstraction or multi-step reasoning that only a very small portion of the sample could handle. Items at the opposite end of the continuum were much more accessible; students tended to answer them correctly at higher rates, which indicates that foundational skills were consistently mastered.

Although this alignment was not perfect, the overall targeting was reasonable. The test captured performance differences for students in the low-to-mid ability ranges, which is typically the range in which diagnostic assessments are expected to be most sensitive. However, the limited spread of highly difficult items means that the instrument offered less information about students performing at the upper end of the continuum. In practical terms, the Wright Map suggests that the instrument functioned effectively in diagnosing the abilities of most students; however, it may benefit from the inclusion of additional high-complexity tasks if the goal is to capture more advanced modelling and reasoning skills. This finding helps frame the next

Department of Mathematics Education, Universitas Muhammadiyah Purwokerto, Purwokerto, Indonesia
p-ISSN 2477-409X, e-ISSN: 2549-9084

analyses, which explore how difficulty patterns vary across PISA indicators, content domains, and grade levels.

Item Difficulty Analysis by the Five PISA Dimensions

To examine where students encountered the greatest difficulty, item measures were examined across several dimensions of the PISA mathematical literacy framework. Table 5 summarizes the overall pattern. Instead of reviewing every value in the table, the focus here is on the key contrasts that help explain the conceptual challenges students experienced.

Table 5. Presents The Mean Logit Difficulty & Proportion of Correct Responses for Each Indicator

Aspect	Mean Measure	% Correct	Min	Max
Indicator				
Mathematical Reasoning	0.13	46.13	-1.20	2.37
Formulating situations mathematically	0.51	38.38	-0.59	2.65
Employing mathematical concepts, facts, and procedures	0.04	50.18	-1.34	1.92
Interpreting, applying, and evaluating mathematical outcomes	-0.68	71.22	-1.59	-0.29
Level				
Level 1c	-0.97	76.75	-1.59	-0.37
Level 1b	-0.33	56.83	-0.91	0.27
Level 1a	-0.22	56.83	-0.94	0.20
Level 2	0.97	27.31	-0.63	2.37
Level 3	0.08	49.82	-1.20	2.65
Level 4	0.34	42.07	-0.35	2.01
Level 5	-0.04	50.18	-0.43	0.99
Level 6	0.21	46.13	-0.57	1.92
Content				
Space and Shape	0.35	42.07	-0.63	2.37
Quantity	-0.60	64.21	-1.34	-0.23
Change and Relationship	0.06	50.18	-1.59	2.65
Uncertainty and Data	0.07	49.82	-0.91	2.01

The four indicators presented noticeably different levels of difficulty. This indicator had the highest mean logit and the lowest accuracy, suggesting that many students found it difficult to translate contextual information into a mathematical representation. By contrast, the interpreting indicator showed the lowest difficulty, and Figure 2 illustrates how consistently this indicator remained below the others. Students appeared more comfortable working with results once the mathematical

structure had already been provided. The employing indicator fell in the middle, pointing to an interesting pattern in which procedural skills were stronger than modelling skills, but not always sufficient to handle more complex tasks.

The researcher analyzed the distribution of item difficulty to map student performance across multiple aspects of mathematical literacy. This evaluation encompasses cognitive process domains, PISA proficiency levels, and mathematical content. The analysis used the mean logit measure and the percentage of correct responses to identify the areas that were most challenging for students. The statistical data regarding mean difficulty and the proportion of correct responses are presented in Table 5.

Table 5 summarizes the difficulty levels based on process, level, and content aspects. The data in the table indicates that "Formulating situations mathematically" has the highest mean logit of 0.51, with the lowest percentage of correct responses at 38.38%. Conversely, "Interpreting, applying, and evaluating mathematical outcomes" is the most accessible aspect for students, with a logit value of -0.68 and a success rate of 71.22%.

Based on the proficiency levels, Level 2 items were identified as the most difficult, with a mean logit of 0.97. Meanwhile, Level 1c items exhibited the lowest level of difficulty, with a mean logit of -0.97 and a correct response rate of 76.75%. Regarding content categories, "Space and Shape" posed the greatest challenge for students, with a mean logit of 0.35. The "Quantity" domain demonstrated the best student performance, with a mean logit of -0.60 and a success rate of 64.21%. Overall, the findings presented in Table 5 demonstrate variations in students' abilities to resolve different dimensions of mathematical literacy. The prevalent difficulty in formulating situations and managing "Space and Shape" content provides critical information for future instructional strategies. This data proves that the instrument effectively maps student strengths and weaknesses in detail.

The researcher presents a data visualization to clarify the differences in difficulty levels across each mathematical literacy process indicator. Through this graph, the comparisons between reasoning, formulating, employing, and interpreting abilities were more clearly observed. This visualization assisted the researcher in identifying performance patterns relative to the applied PISA standards. A comparison of the mean logit values for each process indicator is presented in Figure 1.

Figure 1 illustrates the distribution of mean logit difficulty for the four primary mathematical literacy indicators. The data presented in the graph shows that the

Department of Mathematics Education, Universitas Muhammadiyah Purwokerto, Purwokerto, Indonesia
p-ISSN 2477-409X, e-ISSN: 2549-9084

"Formulating" indicator had the highest logit value of 0.51, indicating the greatest level of difficulty for students. Conversely, the "Interpreting" indicator had the lowest logit value of -0.68. This significant difference suggests that students found it considerably easier to interpret mathematical outcomes than to formulate situations into mathematical forms.

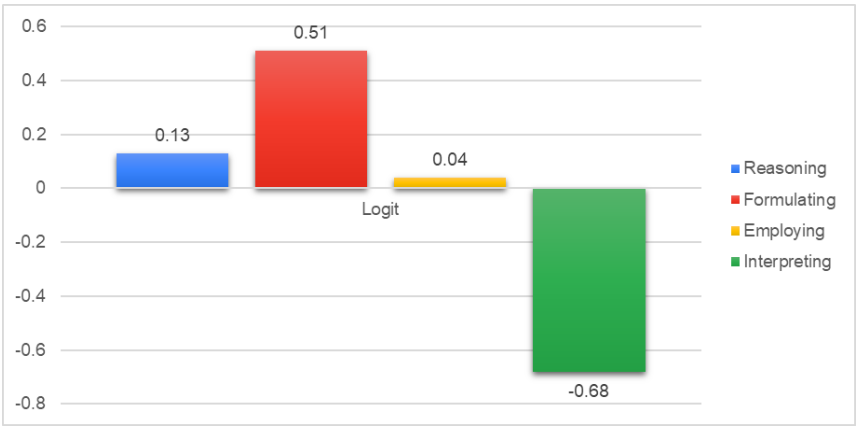

Figure 1. Mean logit difficulty by indicator

The progression of student proficiency across levels followed a pattern that was generally aligned with the expectations of the PISA framework. Based on Figure 1, there was a sharp decline in accuracy once the items began to demand reasoning skills rather than simple recognition. Students performed relatively well on lower-level tasks but showed noticeable difficulty when progressing to assignments that required more than one conceptual step. This phenomenon suggests that higher-level reasoning skills are still developing and require stronger instructional support. Overall, Figure 1 provides visual evidence that formulating and reasoning aspects constitute the primary obstacles for students in mathematical literacy.

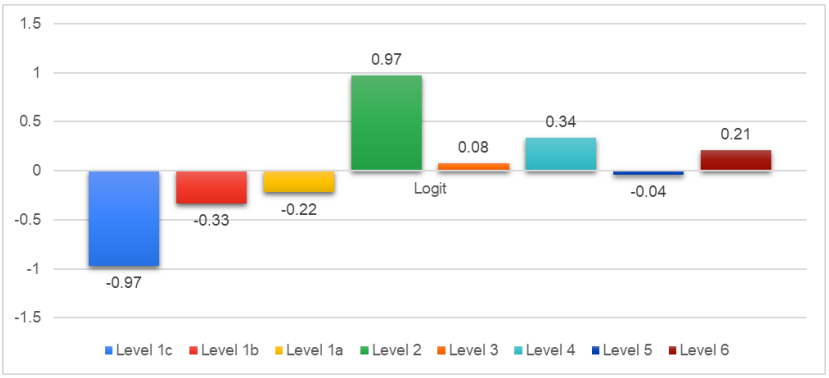

Figure 2. Mean logit difficulty by level

Figure 2 illustrates the distribution of mean logit difficulty based on PISA

mathematical literacy levels. The data presented in the graph shows that Level 1c was the easiest stage for students, with a logit value of -0.97. Conversely, Level 2 was recorded as the stage with the highest average difficulty, with a logic value of 0.97. Although higher proficiency levels are theoretically expected to be more difficult, the findings presented in Figure 2 show fluctuations at the middle and upper levels, such as Level 4 (0.34) and Level 6 (0.21), which may have been influenced by the specific characteristics of the items used.

Differences in performance also appeared across various content and context categories embedded within these levels. Many students struggled with spatial reasoning and geometric structures at higher proficiency levels. In contrast, the Quantity content was the easiest category, reflecting students' familiarity with numerical procedures. Context also played a role in item difficulty, as shown in Figure 2. Societal contexts were the most difficult, followed by scientific and occupational settings, while personal contexts were comparatively easier. This phenomenon suggests that students tend to experience difficulty connecting mathematical reasoning when ideas are embedded in abstract or less familiar real-world situations.

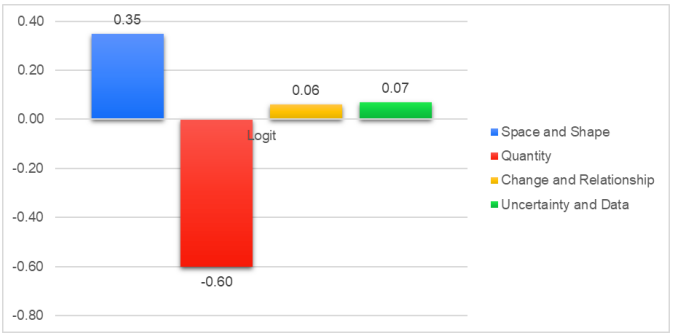

Figure 3. Mean logit difficulty by content category

Figure 3 illustrates the distribution of item difficulty levels across four primary content categories. A closer examination of individual item measures showed that most items fall within the moderate difficulty range. However, the analysis presented indicates that the specific challenge in multi-step reasoning did not lie in the final calculation. Instead, the primary hurdle appeared during the "integrative transition," in which students were required to carry the output of one mathematical procedure forward as the input for the next while maintaining logical consistency.

Based on the data in Figure 3, the content categories of Space and Shape and Change and Relationship demonstrated higher difficulty levels than the other categories. In modelling tasks, the main conceptual obstacle was "structural abstraction," which refers to the ability to isolate critical variables from descriptive text and determine their

Department of Mathematics Education, Universitas Muhammadiyah Purwokerto, Purwokerto, Indonesia
p-ISSN 2477-409X, e-ISSN: 2549-9084

mathematical relationships. Students often failed at the initial stage of defining the problem's structure, leading to a breakdown before any formal operations can be applied. In contrast, a small group of items was considered easy, largely reflecting recognition-based or routine tasks.

A closer look at the individual item measures shows that most items fell within the moderate difficulty range. The specific challenge in multi-step reasoning was not found in the final calculation, but in the "integrative transition" in which students were required to carry the output of one mathematical procedure as the input for the next, while maintaining logical consistency. Furthermore, in modelling tasks, the primary conceptual hurdle was "structural abstraction", the ability to isolate critical variables from descriptive text and determine their mathematical relationships. Students often failed at the initial stage of defining the problem's structure, leading to a breakdown before any formal operations could be applied. A small group of items was considered easy, largely reflecting recognition-based or routine tasks. The spread of difficulty levels indicated that the test captured a range of abilities while still maintaining good alignment with the student cohort. Overall, the patterns across indicators, proficiency levels, content categories, and contextual settings suggest that the main areas of difficulty lie in formulating and spatial tasks. These aspects consistently appeared at the higher end of the difficulty continuum and will be important points of focus in the discussion of implications and curriculum alignment.

The researcher classified the 32 test items according to their logit values to comprehensively map the difficulty distribution of the instrument. This grouping aimed to ensure that the instrument maintained a balance between easy, moderate, and difficult items. Through this classification, the researcher could evaluate whether the instrument effectively covered a range of student proficiency levels, from basic to advanced. A summary of the item difficulty classification is presented in Table 6.

Table 6. Logit Value Item Analysis (N=32)

Difficulty Level (logit)	Item Code (logit)	Total (%)
Easy (Measure <-1.04)	02ECPQ (-1.34), 03ICCC (-1.59), 20R3CC (-1.2)	3 (9.38)
Moderate (-1.04≤ Measure ≤ 1.04)	01FCPQ (-0.59), 04RCSS (-0.37), 05FBSS (0.27), 06EBOU (-0.37), 07IBOU (-0.91), 08RBPQ (-0.29), 09FAOU (0.1), 10EACC (0.2), 11IAPQ (-0.47), 12RAPQ (-0.94), 14E2OU (0.88), 15I2SS (-0.63), 18E3CC (-0.35), 19I3OU (-0.79), 21F4OU (-0.06), 22E4PQ (-0.23), 23I4PQ (-0.35), 25F5SS (0.99), 26E5SS (-0.43), 27I5SS (-0.37), 28R5OU (-0.33), 29F6SS (-0.57), 31I6CC (-0.29), 32R6CC (-0.21)	24 (75)
Difficult (Measure >1.04)	13F2CC (1.25), 16R2SS (2.37), 17F3CC (2.65), 24R4OU (2.01), 30E6SS (1.92)	5 (15.63)

Table 6 presents the categorization of test items into three primary groups based on their logit measures. The data in the table indicate that the majority of the items fall within the moderate category. A total of 24 items, or 75% of the instrument, had logit values between -1.04 and 1.04. This demonstrates that the instrument was highly focused on measuring the average proficiency level of the student group.

Based on Table 6, 3 items (9.38%) were classified as easy, with logit values below -1.04, such as item 03ICCC, which has the lowest value of -1.59. Meanwhile, the difficult category includes 5 items (15.63%) with logit values exceeding 1.04. Item 17F3CC was recorded as the most challenging question, with a logit value of 2.65. The distribution, dominated by moderate-level items, indicates that the instrument is well-aligned with the student population while still providing challenges for high-ability students through more complex items.

Comparative Analysis by Grade Level

The researcher compared students' ability estimates across Grades VII, VIII, and IX to explore the development of mathematical literacy across different levels. This analysis aimed to determine whether there was a linear increase in proficiency as students progressed through their education. By comparing mean ability values across various indicators, levels, and content categories, the researcher could identify developmental trends and performance anomalies at specific grades. A summary of mean student ability by grade level is presented in Table 7.

Table 7 illustrates the distribution of student abilities measured across various aspects of mathematical literacy. Based on the data, Grade VIII students achieved the highest proficiency level with a global mean ability of 0.257. This was followed by Grade VII with a value of 0.020, while Grade IX students scored the lowest at -0.053. This phenomenon indicates a decline in performance for Grade IX, which disrupts the expected upward trend in proficiency. Such patterns are also observed in other literacy-based assessments, where older students may encounter more demanding tasks or experience test fatigue.

A deeper examination of Table 7 highlights the advantage of Grade VIII students in the aspects of Formulating (0.299) and Interpreting (0.295). Both aspects require students to connect contextual information with mathematical structures. In contrast, Grade IX shows values below zero across almost all indicators, suggesting that modelling and reasoning skills at that level have not developed steadily. The proficiency curves across grades run mostly parallel, signaling that the ordering of item difficulty remains stable across all grade levels.

Department of Mathematics Education, Universitas Muhammadiyah Purwokerto, Purwokerto, Indonesia
p-ISSN 2477-409X, e-ISSN: 2549-9084

Table 7. Mean Ability by Grade Levels

Aspect	Ability by Grade Levels		
	VII	VIII	IX
Indicator			
Mathematical Reasoning	-0.075	0.244	-0.076
Formulating situations mathematically	-0.071	0.299	-0.073
Employing mathematical concepts, facts, and procedures	0.020	0.211	-0.028
Interpreting, applying, and evaluating mathematical outcomes	0.185	0.295	-0.034
Level			
Level 1C	0.022	0.218	-0.174
Level 1B	0.002	0.146	-0.010
Level 1A	0.080	0.201	-0.116
Level 2	-0.054	0.176	0.044
Level 3	0.150	0.324	-0.129
Level 4	0.029	0.258	0.002
Level 5	-0.076	0.290	-0.096
Level 6	-0.036	0.316	-0.031
Content			
Space and Shape	-0.074	0.237	-0.019
Quantity	0.140	0.250	-0.099
Change and Relationship	-0.001	0.315	-0.093
Uncertainty and Data	0.006	0.210	-0.047
Global	0.020	0.257	-0.053

Figure 4 illustrates the comparison of abilities among Grade VII, VIII, and IX students across the 32 tested items.

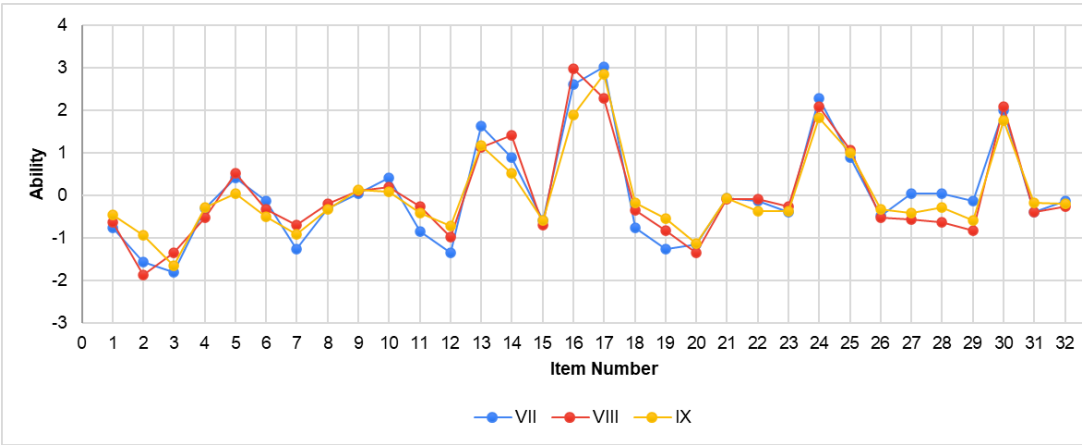

Figure 4. Person DIF plot based on the difference in students' grade level

In general, the Quantity domain remained the easiest for all three grades. Conversely, the Space and Shape and Change and Relationship domains present higher levels of demand. Grade IX students demonstrated a consistent decline in ability across almost

all domains. This finding reinforces the view that conceptual understanding does not always increase with additional years of schooling. Significant growth in mathematical literacy was observed between Grades VII and VIII, but this progress did not continue uniformly into Grade IX. This slight decline suggests that advanced reasoning and modelling skills require targeted instructional support rather than passive exposure through regular instruction. The patterns observed in Figure 1 and Figure 4 serve as crucial considerations in interpreting the diagnostic implications of the Rasch analysis in the discussion section.

The first research question examined which PISA indicators were most conceptually demanding for students. The results clearly positioned Formulating as the most difficult indicator, while Interpreting emerged as the least challenging. This pattern suggests that students struggled more with the initial stages of mathematical modelling, extracting information, determining relationships, and constructing a mathematical representation, than with interpreting results once the structure was known. Several studies reflect this trend from different perspectives. The interpretation of this disparity lies in the "instructional fossilization" of procedural habits; students are accustomed to being given ready-made formulas, making the act of "meaning-making" and structural construction in Formulating the primary barrier, rather than the final reflection. Students with analytic cognitive styles tended to achieve better results in representation tasks, indicating that individual cognitive preferences significantly influence modelling success (Nurfitriani et al., 2024).

From an international perspective, early adolescents in Ireland demonstrate proficiency in interpretation tasks but struggle with problem formulation, a challenge primarily attributed to limited exposure to unstructured problems (Leavy et al., 2023). Conversely, students in Spain show improved performance when modelling tasks are strongly connected to familiar cultural contexts, suggesting that context familiarity can effectively offset structural difficulty (Marín-álvarez et al., 2024). There are several possible interpretations of this disparity. Some relate to instruction: Indonesian classrooms have historically emphasized procedural mastery over exploratory modelling, creating a strong competence in recognition but weaker performance in representation. Others relate to assessment: unfamiliar contexts increase cognitive load, disproportionately affecting the formulation stage. The consistent difficulty observed in the 'Formulating' domain suggests that the challenge extends beyond a mere lack of technical skills; rather, it indicates a conceptual gap in which students struggle to perceive mathematical modeling as a meaning-making process. Consequently, instructional strategies should prioritize repeated exposure to open-ended scenarios that require students to negotiate meaning, construct representations, and justify their structural choices.

p-ISSN 2477-409X, e-ISSN: 2549-9084

The second research question concerned how item difficulty varied across content domains and problem contexts. The results showed that Space and Shape presented the greatest challenge, specifically due to students' difficulties in mental rotation and spatial visualization. These challenges were most evident when students were required to perform spatial scaling and transform two-dimensional representations into three-dimensional mental models, rather than simply applying geometric formulas. While Quantity was the most accessible domain, the complexity of Space and Shape suggests that students struggle to maintain the structural properties of shapes in non-routine contexts. Contextually, societal and scientific settings were noticeably harder than personal contexts. The critical interpretation of this study is that the "Space and Shape" difficulty is an artifact of the curriculum's historical minimization of spatial reasoning in favor of numerical computation. Furthermore, while some literature suggests that Personal contexts are always easier, this research posits that context familiarity is not a static advantage; it can be offset by structural complexity. This helps explain why students may struggle even in familiar settings if the "spatial scaling" or "mental rotation" requirements are high.Geometry tasks modeled after the PISA framework are often perceived as challenging due to a lack of prior exposure to spatial modeling activities (Octaria et al., 2025). This struggle with spatial reasoning is also observed among high-performing East Asian students, suggesting that the difficulty may stem from instructional gaps rather than cultural differences unless visualisation tasks are deliberately incorporated into the curriculum (Xu et al., 2025).

Contextual variation also significantly influences these broader findings. Unfamiliar contexts have been found to impose additional linguistic and situational interpretation demands, which complicate the modeling process (Gilligan-Lee et al., 2022). Conversely, students tend to perform better when mathematical tasks, even complex ones, are directly connected to their out-of-school experiences (Marín-álvarez et al., 2024). These mixed findings suggest that difficulty is shaped by how students relate to the situations presentedrather than the context category itself. Looking at these results through multiple lenses, several interpretations emerge. One is curricular: Indonesian textbooks often minimise geometry and contextual reasoning. Another is experiential: many students rarely encounter mathematics situated in civic or scientific contexts. A third is cognitive: spatial reasoning develops more slowly and requires guided experiences. The pronounced challenges in Space and Shape and societal contexts provide empirical evidence that students require more integrated exposure to tasks combining context, representation, and spatial reasoning, rather than a singular focus on numerical procedures (Ma'ruf et al., 2025).

The third research question explored differences across grade levels. The results

showed a non-linear pattern: Grade VIII students outperformed both Grade VII and Grade IX. This suggests that mathematical literacy does not grow in a straightforward, linear fashion. This finding directly challenges the conventional "linear growth" assumption in developmental psychology and education, which posits that literacy should increase commensurate with years of schooling, as found in the studies of (Lin et al., 2024). This suggests that mathematical literacy does not develop in a straightforward, linear fashion. This trend may be attributed to the increased cognitive load associated with the greater complexity of Grade IX curriculum materials, which often demand higher levels of abstraction without sufficient instructional support for modelling processes. Similar patterns have been observed where improvement slows once students enter an exam-oriented cycle (Octaria et al., 2025). This trend is reflected in observations that literacy outcomes often dip when instructional emphasis shifts from sense-making to procedural rehearsal (Nurfitriani et al., 2024). International evidence further illustrates this variation, as a similar decline in literacy performance has been observed among older Chilean students (Wijayanto et al., 2024). Conversely, research in Korea indicates that modeling competence can improve steadily when supported by deliberate and structured instructional design (Lin et al., 2024).

There are several plausible explanations for why Grade IX might decline. Motivation may decline as students face higher academic demands and testing pressure. Instruction may shift toward procedural preparation for national exams, thereby limiting modelling opportunities. Cognitive load may increase as tasks require modelling fluency that has not been adequately built. The critical position of this research is that the Grade IX decline is not a failure of student ability, but a symptom of "curriculum narrowing." As the focus shifts toward high-stakes exam preparation, the "sense-making" required for mathematical literacy is sacrificed for "procedural rehearsal." The decline observed in Grade IX potentially reveals a structural gap in literacy development: as mathematical content becomes more sophisticated, the lack of sustained modelling opportunities may cause conceptual growth to stall. The peak at Grade VIII may represent a developmental stage in which students have accumulated sufficient procedural fluency to handle intermediate tasks, but they struggle to transition to Grade IX requirements, where the cognitive load of integrating complex contexts and abstract representations becomes significantly higher. The peak at Grade VIII represents a "golden window" where students possess procedural fluency but have not yet been subjected to the restrictive pressures of exam-centered instruction that stifles conceptual development in later years.

Conclusion

The Rasch analysis in this study extends beyond a simple performance audit to provide a comprehensive diagnostic framework of mathematical literacy

p-ISSN 2477-409X, e-ISSN: 2549-9084

development. For the first question, the findings reveal a significant "formulation-gap" where students struggle to initiate the modelling cycle. Crucially, within the multiple-choice format, this difficulty reflects challenges in identifying the correct mathematical structure among competing distractors. In contrast, the Interpreting indicator produced the highest accuracy, suggesting that students manage well when the mathematical structure of a problem is already in place. The second research question addressed differences across content and context, and the results pointed to Space and Shape as the most challenging domain, while tasks requiring basic numerical reasoning were handled much more successfully. Context also made a difference, with items situated in societal and scientific settings proving more demanding than personal situations. The third research question examined developmental patterns, and the results revealed a non-linear progression. Students in Grade VIII demonstrated the strongest performance, whereas students in Grade IX achieved slightly lower scores. The significance of these results lies in identifying "stagnation points" in literacy growth. This indicates that literacy is a fragile competency that may plateau or decline when instructional focus shifts toward high-stakes, procedural exam preparation. These patterns carry practical meaning. They suggest that students have developed some procedural fluency and can interpret mathematical outcomes, but have limited experience building representations from real contexts. Modelling and spatial reasoning appear to require more deliberate instructional support. The contrast between indicators also reflects how strongly classroom practices shape literacy outcomes: tasks that begin with well-defined mathematical forms are more familiar, while tasks requiring students to decide how to structure a situation are far less common.

From a theoretical perspective, the findings reinforce the idea that mathematical literacy is not a singular skill but a combination of representation-building, reasoning, and contextual judgment. The persistent difficulty in Formulating supports arguments that modelling should not be seen as an “advanced topic” but as a fundamental component of literacy development. The differences across contexts also highlight that mathematical literacy is shaped not only by students’ cognitive preparation, but also by their sociocultural exposure.

The study has several limitations that should be acknowledged. The sample was drawn from urban schools, and the findings may not fully represent rural or culturally diverse contexts. The broader implications of this study support the need for a paradigm shift in assessment design. Future research must transcend the limitations of closed-ended items by utilizing qualitative "think-aloud" protocols to capture the invisible processes during formulation. A mixed-methods design might have revealed how students interpret contextual information and why they struggle during the

Formulating stage. The analysis also focused on grade-level differences without exploring instructional practices, which may play an important role in shaping the observed patterns. These limitations open several possibilities for future research. Studies incorporating interviews or think-aloud protocols would help clarify how students navigate each step of the modelling cycle. Intervention studies could examine whether structured modelling tasks or spatial reasoning activities improve performance across indicators. Expanding the sample to include rural and under-resourced schools would help determine whether the same difficulty patterns persist across different learning environments. Finally, a longitudinal design could track how mathematical literacy develops across multiple school years, allowing a closer examination of the non-linear progression observed in this study.

Acknowledgement

We sincerely thank the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemdiktisaintek) for funding this research.

Author's Declaration

Author Contribution	:	Ilham Falani: Conceptualization, Writing - Original Draft, Supervision Syamsir Sainuddin: Writing - Review & Editing, Validation
Funding Statement	:	This research was funded by the Directorate General of Higher Education, Research, and Technology, Republic of Indonesia, under the Program Penelitian Fundamental Riset (PFR) – Applied Research Scheme.
Conflict of Interest	:	The authors declare no conflict of interest.
Additional Information	:	-

References

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Aulia, H., Mandailina, V., Mahsup, M., Syaharuddin, S., Rini, W., & Sahraini, A. (2024). Measurement of mathematical literacy in everyday life context: A case study on high school students reviewed by gender. *AlphaMath: Journal of Mathematics Education*, *10*(1), 49–58. https://doi.org/10.30595/alphamath.v10i1.21083

Bond, T. G., Zi Yan, & Moritz Heene. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Britania Raya: Routledge of the Taylor& Francis Group.

Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, *15*(4), 1–7. https://doi.org/10.1187/cbe.16-04-0148

Gedik Altun, S. D., & Morkoyunlu, Z. (2023). The use of error based activities to improve the mathematization competency. *International Journal of Progressive Education*, *19*(3), 134-148. https://doi.org/10.29329/ijpe.2023.546.8

Gilligan-Lee, K. A., Hawes, Z. C. K., & Mix, K. S. (2022). Spatial thinking as the missing piece in mathematics curricula. *NPJ Science of Learning*, *7*(1), 10-25. https://doi.org/10.1038/s41539-022-00128-9

Hayat, B., Putra, M. D. K., & Suryadi, B. (2020). Comparing item parameter estimates and fit statistics of the rasch model from three different traditions. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(1), 39–50. https://doi.org/10.21831/pep.v24i1.29871

Karlimah, K. (2022). How does Rasch modeling reveal difficulty and suitability level the fraction test question ?. *Jurnal Elemen*, *8*(1), 66–76. https://doi.org/10.29408/jel.v8i1.4170

Kholid, M. N., Rofi'ah, F., Ishartono, N., Waluyo, M., Maharani, S., Swastika, A., Faiziyah, N., & Sari, C. K. (2022). What are students' difficulties in implementing mathematical literacy skills for solving PISA-like problem?. *Journal of Higher Education Theory and Practice*, *22*(2), 181–200. https://doi.org/10.33423/jhetp.v22i2.5057

L'Boy, D., & Nazim Khan, R. (2023). A Rasch-model-based hierarchical framework for statistical literacy and learning. *International Journal of Mathematical Education in Science and Technology*, *54*(9), 1874–1887. https://doi.org/10.1080/0020739X.2023.2261453

Leavy, A., Hourigan, M., & McMahon, K. (2023). Early adolescents' performance on mathematical modelling tasks: An Irish perspective. *Educational Studies in Mathematics*, *112*(3), 455–478. https://doi.org/10.1007/s10649-022-10196-z

Lee, D., & Yeo, S. (2022). Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Computers & Education*, *191*, 104646. https://doi.org/10.1016/j.compedu.2022.104646

Lin, T. J., Buckley, J., Gumaelius, L., & Ampadu, E. (2024). Locating the potential development of spatial ability in the Swedish national curriculum. *Heliyon*, *10*(19), e38356. https://doi.org/10.1016/j.heliyon.2024.e38356

Ma'ruf, A. H., Triyono, A., Riaseh, A. G., Nuary, R. H., Permatasari, N., & Saleh, R. R. M. (2024). Correlation between mathematical literacy abilities and students' mastery of problem solving abilities. *AlphaMath: Journal of Mathematics Education*, *10*(2), 295–308. https://doi.org/10.30595/alphamath.v11i2.28590

Ma'ruf, A. H., Triyono, A., Fauziah, N. R., Warohmah, M., Nurimani, N., & Kusuma, A. P. (2025). Mapping students' mathematical literacy skills in basic geometry: A study based on Stacey and Turner's indicators. *AlphaMath: Journal of Mathematics Education*, *11*(2), 139–258. https://doi.org/10.30595/alphamath.v10i2.24175

Marín-álvarez, F., Flores-prado, L., Figueroa, O., Polo, P., Varela, J. J., & Muñoz-reyes,

J. A. (2024). Quantitative evaluation of a theoretical-conceptual model based on affective and socio-behavioral dimensions to explain the academic performance of mathematics students. *Frontiers in Psychology*, *August*, 1–11. https://doi.org/10.3389/fpsyg.2024.1372427

Ministry of Education, Culture, Research, and T. of I. (2024). *Regulation of the Minister of Education, Culture, Research, and Technology Number 12 of 2024 on the curriculum for early childhood education, primary education, and secondary education*. Jakarta: Kemdiktisaintek.

Nurfitriani, Y., Yusuf, Y., & Koswara, U. (2024). Mathematical literacy ability of students from a cognitive style perspective on rational. *Edumatsains*, *9*(1), 319–331. https://doi.org/10.33541/edumatsains.v9i1.5985

Nurjanah, A., Saputra, D. C., Garuda, J., & Sleman, K. (2023). Mathematical literacy skills : A survey in secondary schools. *Jambura J. Math. Educ.*, *4*(1), 35–49. https://doi.org/10.34312/jmathedu.v4i1.18776

Octaria, D., Zulkardi, Z., Putri, R. I. I., & Hiltrimartin, C. (2025). Spatial literacy in geometry learning : A systematic literature review. *Indiktika : Jurnal Inovasi Pendidikan Matematika*, *7*(1), 316–324. https://doi.org/10.31851/indiktika.v7i1.17038

OECD. (2023a). *PISA 2022 assessment and analytical framework* (PISA). Paris: OECD Publishing.

OECD. (2023b). *PISA 2022 results (volume i and ii) - country notes: Indonesia*. Paris: OECD Publishing.

Oliva, J. M., & Blanco, Á. (2023). Rasch analysis and validity of the construct understanding of the nature of models in Spanish-speaking students. *European Journal of Science and Mathematics Education*, *11*(2), 344–359. https://doi.org/10.30935/scimath/12651

Ramalisa, Y., Falani, I., & Pasaribu, F. T. (2023). Rasch analysis in developing Jambi culture-based ethnomathematics test for prospective mathematics teachers. *JRAMathEdu (Journal of Research and Advances in Mathematics Education)*, *8*(4), 243-257. https://doi.org/10.23917/jramathedu.v8i4.2921

Risdiyanti, I., Zulkardi, Putri, R. I. I., & Prahmana, R. C. I. (2024). Mathematical literacy learning environment for inclusive education teachers: A framework. *Journal on Mathematics Education*, *15*(3), 1003–1026. https://doi.org/10.22342/jme.v15i3.pp1003-1026

Rusyid, H. K., Suryadi, D., Herman, T., Adnan, M., Lutfi, A., & Mukhibin, A. (2024). Rasch modelling approach to measure the quality of algebraic thinking test item for junior high school students. *Beta: Jurnal Tadris Matematika*, *17*(1), 44–58. https://doi.org/10.20414/betajtm.v17i1.652

Susanta, A., Sumardi, H., Susanto, E., & Retnawati, H. (2023). Mathematics literacy task on number pattern using bengkulu context for junior high school students.

Journal on Mathematics Education, *14*(1), 85–102. https://doi.org/10.22342/jme.v14i1.pp85-102

Truong, Q. C., Gattis, M., Barber, C. C., Middlemiss, W., Au, T., & Medvedev, O. N. (2024). Applying Rasch methodology to examine and enhance precision of the baby care questionnaire. *Journal of Child and Family Studies*, *33*(1), 166–178. https://doi.org/10.1007/s10826-023-02772-0

Wijayanto, Z., Sukestiyarno, S., Wijayanti, K., & Pujiastuti, E. (2024). Analysis of mathematical literacy through the lens of students' spatial geometry aptitude. *Jurnal Cakrawala Pendidikan*, *43*(3), 746–755. https://doi.org/10.21831/cp.v43i3.65013

Wind, S. A., & Hua, C. (2022a). Examining the principles of invariant measurement in Rasch measurement theory. *Journal of Applied Measurement*, *23*(2), 180–195. https://doi.org/10.1201/9781003174660-6

Wind, S. A., & Hua, C. (2022b). Many Facet Rasch Model. In *Rasch Measurement Theory Analysis in R* (pp. 195–278). Bookdown. org,[Epub]. https://doi.org/10.1201/9781003174660-6

Xu, T., Sun, S., & Kong, Q. (2025). Spatial reasoning and its contribution to mathematical performance across different content domains: Evidence from Chinese students. *Journal of Intelligence*, *13*(4), 41-56. https://doi.org/10.3390/jintelligence13040041

Department of Mathematics Education, Universitas Muhammadiyah Purwokerto, Purwokerto, Indonesia
p-ISSN 2477-409X, e-ISSN: 2549-9084