Jurnal Pendidikan MIPA
e-ISSN: 2685-5488 | p-ISSN: 1411-2531
URL: https://jpmipa.fkip.unila.ac.id/index.php/jpmipa
Jounal's DOI: http://doi.org/10.23960/jpmipa

Vol. 27, No. 1, 2026
Article ID: 1160
Pages: 357-382

Psychometric Evaluation of Culturally Responsive Numeracy and Problem-Solving Assessment: A Rasch Modeling Approach

Indah Suciati[1,4*], Daud Karel Walanda[2], Mustamin Idris[3], Mery Napitupulu[2], Anggraini[3], Rita Lefrida[3], & Richard Albert Walanda[5]

[1]Doctoral Program in Science Education, Universitas Tadulako, Indonesia
[2]Department of Chemical Education, Universitas Tadulako, Indonesia
[3]Department of Mathematics Education, Universitas Tadulako, Indonesia
[4]Department of Mathematics Education, Universitas Alkhairaat, Indonesia
[5]Department of Computer Science, University of Newcastle, Australia

*Corresponding email: ndahmath@gmail.com

Received: 25 December 2025 *Accepted: 02 February 2026* *Published: 06 March 2026*

Abstract: The quality of measurement in educational research is highly dependent on the validity and reliability of the instruments used. However, in practice, many numeracy literacy and problem-solving instruments have not undergone adequate psychometric testing, particularly those that use cultural contexts as question stimuli. This study aims to develop and evaluate instruments for numeracy literacy and problem-solving in sequences and series, grounded in cultural context/local wisdom, using a research and development approach and Rasch modeling. The instrument's content validity was assessed through expert review by learning evaluation experts, mathematics education lecturers, and cultural experts (traditional councils). The analysis was conducted using the average validator score, Aiken's V coefficient, and percentage agreement to ensure the appropriateness of the substance, clarity of language, instrument appearance, and the appropriate integration of cultural context. Empirical trials were conducted in small groups with 71 high school students in grade X with 15-17 years in three regions in Indonesia: Palu City, Donggala Regency, and Sigi Regency, who were selected purposively. The test data were analyzed using Rasch modeling to evaluate item suitability, difficulty level, reliability, separation index, unidimensionality, and potential bias through DIF analysis. The analysis results showed that the developed instrument had very high content validity, strong person and item reliability, and that most items fit the Rasch model. However, the distribution of item difficulty did not fully cover the extreme range of respondents' abilities, and indications of DIF were found in some items in one area, indicating the need for further refinement. Thus, the developed numeracy literacy and problem-solving instrument was deemed suitable based on initial findings at the development stage; however, it still requires further revision and testing before being widely used in a more diverse population.

Keywords: numeracy literacy, problem-solving, rasch model, reliability, validity.

Article's DOI: https://doi.org/10.23960/jpmipa.v27i1.pp357-382

INTRODUCTION

Education in today's era of globalization demands the development of human resources with critical, creative, and adaptive thinking skills. Numeracy, literacy, and problem-solving skills are essential competencies in the 21st century, enabling students to navigate everyday life effectively (Adams et al., 2020; Amalina & Vidákovich, 2023; Charoentham et al., 2025; Nisa et al., 2024; Ramadhani et al., 2025). These two abilities are not only related to the mastery of mathematical concepts and procedures but also encompass the capacity to reason, solve contextual problems, and make data- and information-based decisions (Amalina & Vidákovich, 2023; Nahdi et al., 2020; Syifa et

al., 2024). Therefore, accurate measurement of numeracy literacy and mathematical problem-solving is an important prerequisite in educational research, particularly for developing and evaluating learning innovations and pedagogical interventions (Rahman et al., 2016; Suciati et al., 2020).

In educational research, test instruments are the primary factor determining the quality of research conclusions (Hellstrand et al., 2020; Ramadhani et al., 2025; Sari et al., 2024; Swarni et al., 2024). Invalid and unreliable instruments can produce biased interpretations of students' abilities, thereby obscuring the effectiveness of a learning model or educational intervention (Ariffin et al., 2010; Avinç & Doðan, 2024; Haeruddin et al., 2019; Nisa et al., 2024). Therefore, the development of numeracy literacy and problem-solving instruments must be accompanied by comprehensive psychometric quality testing, especially regarding the validity and reliability of the measurement (Gusmanida et al., 2024; Nisa et al., 2024; Ramadhani et al., 2025; Suciati et al., 2020).

Various studies show that many test instruments used in educational research still rely on classical statistical approaches, such as assessing item validity and reliability based solely on alpha coefficients, without considering item characteristics in greater depth (Olkun et al., 2016; Reffiane et al., 2021; Zainil et al., 2024). This condition may cause the instrument to fail to represent the theoretical construct being measured and to be less sensitive to differences in student abilities (Ariffin et al., 2010). In addition, instruments have often not been tested for item functional stability across different respondent groups, thereby risking measurement bias that could affect the fairness and accuracy of research results (Ariffin et al., 2010; Husain & Aziz, 2022).

Modern psychometric approaches, particularly the Rasch Model, offer a more objective and robust analytical framework for evaluating instrument quality (Sari et al., 2024; Swarni et al., 2024). The Rasch model provides more in-depth empirical information on individual item characteristics, assesses item fit to the model, maps the distribution of item difficulty levels and respondent abilities onto the same scale, and more accurately estimates the instrument's reliability and separation index. Furthermore, Rasch modeling enables the detection of potential measurement bias through differential item functioning analysis, thereby supporting the principles of fairness and accuracy in educational assessment (Ariffin et al., 2010; Avinç & Doðan, 2024; Raof et al., 2021; Syifa et al., 2024).

Recent studies at the global level show a shift in the development of numeracy literacy and problem-solving instruments from measuring procedural abilities to assessments that emphasize mathematical reasoning in meaningful real-life contexts (Adams et al., 2020; Chen et al., 2020; Jackson, 2022; OECD, 2012; Platas et al., 2014). This direction aligns with the principle of culturally responsive assessment, which views cultural context as a means of enhancing construct validity and the relevance of assessment instruments (Evans & Taylor, 2025; Ladson-Billings, 1995; Walker et al., 2023). At the national level, several studies have integrated local cultural elements, such as musical instruments, buildings, and social practices, into the development of numeracy instruments to bridge the gap between school mathematics and students' learning experiences (Andrian et al., 2025; Anriana et al., 2023; Nisa et al., 2024; Pratama & Yelken, 2024). However, most studies still focus on context development and content validation, while testing instrument quality using modern psychometrics, particularly Rasch, remains limited. This situation indicates the need to develop culturally grounded numeracy literacy and problem-solving instruments that are not only contextual but also possess strong, stable psychometric properties acro respondent groups.

Recent research in educational assessment indicates a growing use of modern psychometric approaches to ensure instrument quality (Swarni et al., 2024). Several studies have used the Rasch model to analyze the quality of literacy and problem-solving instruments separately (Amalina & Vidákovich, 2022; Nisa et al., 2024; Utami et al., 2025). However, studies that systematically develop instruments grounded in theory, verify content validity through expert assessment, and comprehensively evaluate instrument validity and reliability using the Rasch Model remain relatively limited, particularly in assessing high school students' numeracy literacy and problem-solving abilities.

The novelty of this research lies in the systematic integration of theoretical studies, multidisciplinary content validity, and Rasch model-based psychometric analysis within a single framework for evaluating numeracy literacy and problem-solving instruments. This differs from previous research, which generally relied on classical analysis (Hellstrand et al., 2020; Olkun et al., 2016; Reffiane et al., 2021; Zainil et al., 2024), or used Rasch partially and separately from the theoretical basis and expert validation (Ramadhani et al., 2025; Syifa et al., 2024), This study positions these three approaches as complementary and sequential processes in psychometric decision-making. This integration enables a consistent link among theoretical constructs, instrument content representation, and empirical evidence of item quality, resulting in more objective measurements. Furthermore, the context of local wisdom/culture is treated as a stimulus variable within the items, rather than as a measurement construct, thereby maintaining the unidimensionality and objectivity of the Rasch model. Cross-regional item-function testing through Differential Item Function (DIF) analysis provides initial empirical evidence regarding the fairness and equivalence of the measurements (Debelak et al., 2022; Raof et al., 2021).

Given prior research and the limitations of empirical studies, this study focuses on analyzing the quality of numeracy literacy instruments and problem-solving using the Rasch Model. Specifically, this research is directed to answer the following research questions:

RQ: How is the validity and reliability of numeracy literacy and problem-solving instruments reviewed based on the Rash Model analysis?

By answering these questions, this research is expected to produce an instrument with strong psychometric properties and suitability for use in educational research and practice.

METHOD

Participants

The subjects in this study were 71 senior high school (SMA) students from three educational units, namely SMAN 3 Balaesang (Donggala Regency), SMAN 8 Palu (Palu City), and SMAN 8 Sigi (Sigi Regency). All respondents were grade 10 students aged 15 to 17 years. By gender, there were 28 male and 43 female students. The selection of the three schools was conducted using purposive sampling, with regional representation and the characteristics of schools with low numeracy literacy achievement levels as criteria, based on Education Report Card data. The number of respondents was deemed adequate for a pilot study (small group) in the initial development study to evaluate item quality and measurement parameter stability using Rasch modeling.

The justification for using the number of samples in Rasch model analysis is also supported by Linacre's findings, summarized by Winckel et al. (2022), which states that the stability of item calibration within ±1 logit can be achieved with a relatively small sample size (N = 30), especially in exploratory studies with good targeting between respondent ability and item difficulty. For polytomous scales, a sample size of

approximately 50 respondents is recommended to obtain more stable parameter estimates with high confidence. In line with this recommendation, previous studies have shown that Rasch analysis can be performed adequately with samples of fewer than 100 (Mohamad et al., 2015; Raof et al., 2021; Syifa et al., 2024). Thus, the involvement of 71 respondents in this study exceeded the recommended minimum limit and was therefore considered sufficient to obtain relatively stable item parameter estimates. Therefore, the Rasch analysis results in this study are interpreted as preliminary findings at the instrument development stage and are not intended for generalization to a wider population.

Research Design

This study employed a research and development (R&D) approach to develop and evaluate numeracy literacy and problem-solving instruments. The primary focus of the study is on testing the instrument's feasibility as a measuring tool, particularly in terms of validity and reliability, using a modern psychometric approach, namely the Rasch Model, to ensure that the instrument has objective, consistent, and bias-free measurement qualities (Avinç & Doðan, 2024; Husain & Aziz, 2022; Lin et al., 2020; Raof et al., 2021; Suciati et al., 2020; Swarni et al., 2024). Therefore, this research is not intended to test the effectiveness of a learning model, but rather to evaluate the instrument's suitability as a measure of numeracy literacy and problem-solving skills in the context of educational research (Sari et al., 2024; Swarni et al., 2024). This research can contribute to the literature on instrument validation in the context of local culture- and wisdom-based mathematics education in developing countries, particularly in Indonesia.

Research Procedures

The research procedure was conducted in a systematic, staged manner, as shown in Figure 1. The initial stage began with a theoretical study to develop indicators of numeracy literacy and problem-solving skills, informed by the literature and curriculum. Based on these indicators, an instrument grid was prepared as a reference for developing test items. The next stage was the preparation of test items, followed by content validity through expert assessment to ensure the suitability of the substance, clarity of language, and representativeness of the construct indicators (Ramadhani et al., 2025). The revised instrument, informed by expert input, was then piloted in a small-group setting. The pilot data were analyzed using Rasch modeling to evaluate the instrument's psychometric quality, including item characteristics and overall measurement quality. The results of the Rasch analysis were used to revise the instrument items, thereby ensuring a more stable and suitable instrument.

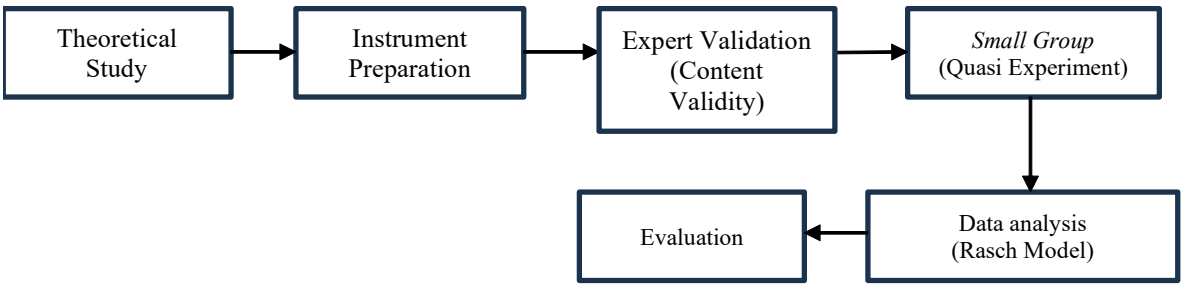

Figure 1. Research procedures

Instrument

The research instrument consisted of an expert-validated questionnaire and a written test, developed systematically to measure numeracy literacy and problem-solving skills in the context of sequences and series.

An expert validation questionnaire was used to assess the content validity of the instrument, which included aspects of (1) content suitability, (2) language clarity, (3) appearance (presentation), and (4) integration of cultural context/local wisdom. The questionnaire was compiled using a 1-5 Likert scale and validated by 1 validator in the field of learning evaluation, 2 validators in the field of mathematics education, and 1 validator in the field of culture (traditional council). However, the scores from the traditional council were not combined with those from the other three validators because the traditional council focused solely on cultural context/local wisdom, which were included in the stimulus questions. The assessment results were analyzed using the average validator score, Aiken's V coefficient, and percentage agreement to ensure content accuracy and inter-validator consistency.

A written test was administered to assess the instrument's empirical validity and reliability using Rasch modeling. The written test consists of two types: a numeracy-literacy instrument and a problem-solving instrument. The numeracy literacy instrument comprises 11 Complex Multiple-Choice questions. The development of numeracy literacy indicators is based on the national numeracy literacy framework (Directorate of Senior High School, 2021; Ministry of Education, Culture, Research, and Technology, 2023), which emphasizes the ability to understand, use, and reason about mathematical concepts in real-world contexts.

Table 1. Numeracy literacy ability test grid

No.	Learning Outcome Indicators	Cognitive Level	Question Format	Indicator	Question Number
1.	Explains the meaning of arithmetic and geometric sequences and series, as well as their elements (a, b, r, Un, and Sn)	*Knowing*	Complex Multiple Choice	Identifying sequence elements from explicitly presented narratives or number patterns	1, 2, 7, 8
2.	Determine the nth term of arithmetic and geometric sequences with the values of a, b, and r.	*Applying*	Complex Multiple Choice	Using the formula Un = a + (n – 1) b to solve contextual problems	3, 9
3.	Calculating the sum of the first n terms of arithmetic and geometric series (Sn) in real situations, including infinite geometric series.	*Applying*	Complex Multiple Choice	Solving local cultural problems	4, 10
4.	Constructing models of arithmetic and geometric sequences and series from local cultural problems, including infinite geometric series.	*Reasoning*	Complex Multiple Choice	Organizing information, constructing number patterns, and determining sequence formulas in a cultural context	5, 6, 11

The cultural context/local wisdom is used as a question stimulus without being positioned as the construct being measured. Example of a numeracy literacy (Complex Multiple Choice):

Level	5 (Grade 10)
Domain	Algebra
Subdomain	Relations and Functions (including Number Patterns)
Context	Social Culture (Donggala Woven Cloth)
Competence	Understanding Arithmetic Sequences and Series
Cognitive Level	Knowing
Question Format	Complex Multiple Choice

Donggala woven cloth is a cultural heritage of Central Sulawesi, traditionally produced on a non-mechanical loom (ATBM). One of its distinctive motifs features geometric lines arranged in stages across each row of the fabric. Each row shows a regular increase in the number of motifs, reflecting the mathematical regularity that is an integral part of this traditional art. One craftsman recorded the number of line motifs created in the first few rows of weaving as follows:

Row No.	Number of Motifs
1	4
2	7
3	10
4	13
5	16

If the pattern continues consistently until line 10, select all of the following statements that are correct!

A. The first term (a) of the series pattern is 4.
B. The difference (b) between the terms of the pattern is 3.
C. The 10th term of the pattern is 31.
D. The general formula for the n-th term of the sequence is $U_n = 3n + 1$.
E. The value 22 is in the 9th term.

Figure 2. Example of question in numeracy literacy instrument

The problem-solving instrument is made in the form of a description of 6 questions, which are developed based on the problem-solving stages of Polya (1957), which also refers to the cultural context/local wisdom as a stimulus for questions to bring mathematical problems closer to real-life experiences, as well as formulated ability indicators. The problem-solving instrument grid is presented in Figure 3. Example of problem-solving questions (Description):

Table 2. Problem-Solving ability test grid

No.	Indicator	Cognitive Level	Question Format	Question Number
1.	Explain the concepts and elements of arithmetic and geometric sequences and series based on the information provided.	C2 (Knowing)	Description	1, 4

2.	Determining the nth term and the difference of an arithmetic and geometric sequence and series through a cultural context	C3 (Applying)	Description	2, 5
3.	Develop models and solve local cultural contextual problems using arithmetic and geometric sequences and series	C4 (Reasoning)	Description	3, 6

Topic	Arithmetic Sequences and Series
Context	Social Culture (Traditional Musical Instruments - *Kakula*)
Indicator	Students can construct models and solve local cultural contextual problems using arithmetic sequences and series.
Cognitive Level	C4 (Reasoning)
Question Format	Description

Kakula is a traditional musical instrument of the *Kaili* people, played by striking it. *Kakulas are shaped like metal gongs of varying sizes and are used in various traditional ceremonies, such as welcoming honored guests and in* cultural rituals. In a *Kakula* set, there are several gongs of varying diameters, arranged from largest to smallest to produce notes from low to high. A metal craftsman is making a *Kakula* set comprising 7 gongs of varying diameters, arranged in an orderly sequence to produce a musical scale. It is known that:

- The first gong has a diameter of 60 cm,
- Second gong: 56 cm
- Third gong: 52 cm, and so on

Question:
Construct a mathematical model to calculate the total diameter of all the gongs in a set of Kakula.

Figure 3. Example of problem solving question

Data Analysis

Several techniques were used in analyzing the instrument's content and empirical validity data. Content validity was assessed to ensure the instrument's suitability for measuring numeracy literacy and problem-solving abilities. The content validation process involved experts assessing the instrument's substance, construction, and language. The instrument's suitability was generally determined from the average score of the validator's assessment results, which were then categorized (Table 3). Furthermore, content validity at the item level was assessed using Aiken's V index to quantify expert agreement for each item. The obtained coefficient values were interpreted using the categories in Table 4 (Amalina & Vidákovich, 2022; Mohamad et al., 2015; Reffiane et al., 2021). To ensure consistency of assessments across validators, a percentage agreement analysis was conducted to quantify the level of agreement between assessors. An instrument is considered reliable if it has an R e" 75% or 75% of the validator's average score in the valid category (Asy'ari et al., 2018; Rahmawati & Trimulyono, 2021; Schouten, 1985).

Meanwhile, for empirical validity, the Rasch modeling approach is used. Rasch modeling is used to assess item fit, item difficulty, reliability, separation index, and potential item bias via Differential Item Functioning (DIF) analysis in Winsteps version 5.10.1.0. The analysis focuses on several main indicators, which include (Nisa et al., 2024; Raof et al., 2021): 1) Item Fit through Infit and Outfit Mean Square (MNSQ) values, Z-standardized (ZSTD), and point measure correlation; 2) Distribution of item difficulty levels and respondent abilities through Wright Map; 3) Reliability and separation index for both person and item aspects; 4) unidimensionality test to ensure that the instrument measures one main construct; and 5) Differential Item Function (DIF) analysis to detect potential item bias based on regional groups. The results of the analysis are then categorized based on the acceptance criteria that apply in each approach presented in Table 3, so that conclusions are obtained regarding the suitability of the instrument in measuring numeracy literacy and problem-solving skills in a valid and reliable manner as a basis for making item revision decisions.

Table 3. Interpretation criteria for average validator score results

Ranges	Categories
$4.20 < X \leq 5.00$	Very Valid
$3.40 < X \leq 4.20$	Valid
$2.60 < X \leq 3.40$	Quite Valid
$1.80 < X \leq 2.60$	Less Valid
$1.00 \leq X \leq 1.80$	Invalid

Table 4. Validation results interpretation criteria

Ranges	Categories
$V \geq 0.81$	Very Valid
$0.60 < V \leq 0.80$	Valid
$0.40 < V \leq 0.60$	Less Valid
$V \leq 0.40$	Invalid

Table 5. Rasch modeling indicators

Indicators	Informations
Wright Map	If the left side is substantially higher than the right side, the question is too easy. If the right side is far above the left side, the question is too difficult.
Item Fit	• Infit & Outfit MNSQ (Mean Square): $0.5 \leq MNSQ \leq 1.5$ If MNSQ < 0.5: The items are too predictable (redundant) If MNSQ > 1.5: The items are too random (misfit) • Infit & Outfit ZSTD (Z-Standarized): $-2.0 \leq ZSTD \leq +2.0$ • Point Measure Correlation (PT-Measure Corr): 0,4 ≤ PT-Measure Corr ≤ 0.85 (Items in line with the construct)
Reliability & Separation	• Person Reability: > 0.80 (Good) > 0.90 (Special) • Item Reliability: > 0.90 (Very good) • Separation Index (G): G > 2 (Good)
DIF	If the Prob. If the value of DIF < 0.05, the item is considered biased and should be corrected or removed.
Undimensionality	• Raw Variance Explained by Measure: ≥ 20% • Unexplained Variance in 1st Contrast: < 15% or eigenvalue < 2.0.

Sources: (Avinç & Doğan, 2024; Debelak et al., 2022; Hu & Bentler, 1999; Kline, 2016; Raof et al., 2021)

■ RESULT AND DISCUSSION

The research results presented focus on the development stages of instruments for numeracy literacy and problem-solving skills. The results are presented systematically, following the research procedures: a theoretical review and

indicator formulation; instrument development; content validation through expert assessment; an instrument pilot study with small groups; and psychometric quality analysis using Rasch modeling.

Theoretical Study

The initial stage of the research focused on a theoretical study of high school students' numeracy literacy and problem-solving skills. This study covered four main aspects: ethnomathematics, numeracy literacy, problem-solving, and the characteristics of sequences and series, which were used as the measured content, as presented in Table 6.

The gap between mathematics instruction in schools and students' needs is further reinforced by the results of semi-structured interviews with mathematics teachers from three public schools in Central Sulawesi: SMAN 3 Balaesang, SMAN 8 Palu, and SMAN 8 Sigi. The mathematics teachers reported that students had difficulty understanding sequences and series, particularly in grasping the concepts and applying the correct formulas. This is supported by the results of the analysis of student errors in the sequences and series material presented in Table 7 below.

Table 6. Literature studies

No	Aspects	Findings	Implications for Instrument Development
1	Ethnomathematics Theory	Ethnomathematics holds that every culture has its own mathematical system that can serve as a source of learning (Anriana et al., 2023; Pratama & Yelken, 2024; Rosa et al., 2016).	The integration of local culture/wisdom into mathematics learning can increase the material's relevance and foster students' cultural identity.
2	Numeracy Literacy	Numeracy literacy is not just about counting, but includes the skills of reasoning, analyzing, and interpreting mathematical information in real contexts (Ministry of Education, Culture, Research, and Technology, 2023; OECD, 2023; Setiyani et al., 2024; Umbara & Suryadi, 2019).	Local culture/wisdom can be a strong context for building a more applicable and contextual understanding of numeracy.
3	Problem-Solving	Problem-solving includes recognizing contextual problems, developing strategies, and reflecting on solutions (Lam et al., 2011; Mayer, 1998; Polya, 1957).	Culturally based contextual questions encourage students to think critically and solve problems in contexts with which they are familiar.
4	Sequences and Series	The material on sequences and series teaches the concepts of regularity, patterns, and generalization; however, it is often taught abstractly (Hardiyanti et al., 2022; Zebua et al., 2020).	Cultural contexts, such as weaving motifs, traditional building structures, or traditional rhythms, can be used to connect abstract concepts to students' lived experiences.

Table 7. Analysis of student errors on the topic of sequences and series

No.	Type of Errors	Findings	Example Errors
1	Conceptual Error	• Students misunderstand the definition of sequences and series	• Students state that 2 + 4 + 6 + 8 is a sequence, not a series.

		(students assume that sequences are the same as series) • Students misunderstand the structure of the *n*-th term formula	• Students do not understand the structure of the formula $U_n = a + (n - 1)\, b$.
2	Principle Error	• Students are unable to use formulas correctly. • Students are unable to use the rules of number operations correctly.	• Students use the formula for the *n*-th term of an arithmetic sequence to calculate the sum of *n* terms of an arithmetic sequence. • Students do not yet understand the rules of number operations (positive/negative signs) in the context of sequences and series.
3	Procedural Error	Students make mistakes in the solution steps.	• Students incorrectly substitute known values in the arithmetic series formula. • Students make mistakes in applying algebraic operations when substituting $(n - 1)\, b$.
4	Factual Error	Students forget basic facts about integer arithmetic operations	The student miscalculated the value of 6 x (-4) = 2
5	Operation Errors	Students make mistakes in integer arithmetic operations	Student incorrectly subtracts 120 – 24 to 100
6	Conclusion Drawing Error	• Students draw wrong conclusions from the final answer. • Students are confused or wrong in concluding the final result, especially when it involves information from more than one term.	• Students conclude the results of arithmetic series with arithmetic sequences. • Students are confused about forming equations, for example, $U_5 = 10$ and $U_2 = -5$. • Students are unable to conclude relationships, for example $U_2 = 8$ dan $U_5 + U_3 = 32$.

Curricularly, Phase E Learning Outcomes emphasize students' abilities to generalize quantitative patterns and relationships and apply them in various contexts. This demonstrates a strong alignment between the objectives of the National Curriculum and the Ethnomathematics approach. Therefore, the development of numeracy literacy and problem-solving instruments for sequences and series is designed to integrate cultural contexts/local wisdom into problem situations, while also strengthening cultural identity and developing the profile of Pancasila students, particularly those with critical, creative, and global perspectives. Table 8 provides a more complete explanation.

Instrument Preparation

The instrument development stage builds on the previously conducted theoretical study. It produces a measurement instrument that assesses

Table 8. Mapping of CP, TP, essential material, and cultural context and local wisdom in the sequence and series material

Curriculum Components	Descriptions	Integration of Cultural Context/Local Wisdom
Learning Outcomes (CP)	At the end of phase E, students can generalize the properties of operations with numbers with powers	Weaving Patterns, Rattan Crafts, Stair Structures in Traditional Houses, Traditional Games (such as *nobangan* and *noventili*),

	(exponents), as well as use sequences and series (arithmetic and geometric) in single interest and compound interest.	cooking oil production and fried onion production, determining auspicious days, regional literature (vaino), and the impacts of gold and sand mining.
Learning objectives	Students can recognize, formulate, and solve sequence and series problems, and apply these concepts to single- and compound-interest problems.	Presentation of contextual questions based on cultural activities such as the process of making cooking oil, rattan crafts, traditional games, musical instruments, and so on.
Essential Material	• Arithmetic sequences and series • Geometric sequences and series • Single interest and compound interest	Visualization of motifs; regional literature (Vaino); traditional games; local crafts and products; traditional building structures; determination of auspicious days; and the impact of regional mining, which can be used as a modeling context for rows and series.
Pancasila Student Profile	Critical, independent, creative reasoning, global diversity through the introduction of mathematical patterns in the practice of socio-cultural life/local wisdom.	Local culture and wisdom serve as sources of values, social connectedness, and the strengthening of students' cultural identity.

high school students' numeracy literacy and problem-solving abilities. The indicators are formulated based on theories of numeracy literacy and problem-solving, the characteristics of sequences and series materials, and an ethnomathematics approach, and are translated into operational and contextual test items. Instrument development is oriented not only toward the accuracy of measuring cognitive aspects but also toward the relevance of cultural context/local wisdom as real-world problems. The integration of culture/local wisdom into the instrument is carried out systematically by using cultural practices, cultural products, and other elements as the background for mathematical problems. This approach is expected to enhance the meaningfulness of the questions, encourage contextual reasoning, and assess students' ability to apply concepts of sequences and series in real-world contexts. The results of this stage yield prototype-1, which is ready for expert validation and limited trials. The instrument development process is presented in Figure 4 below.

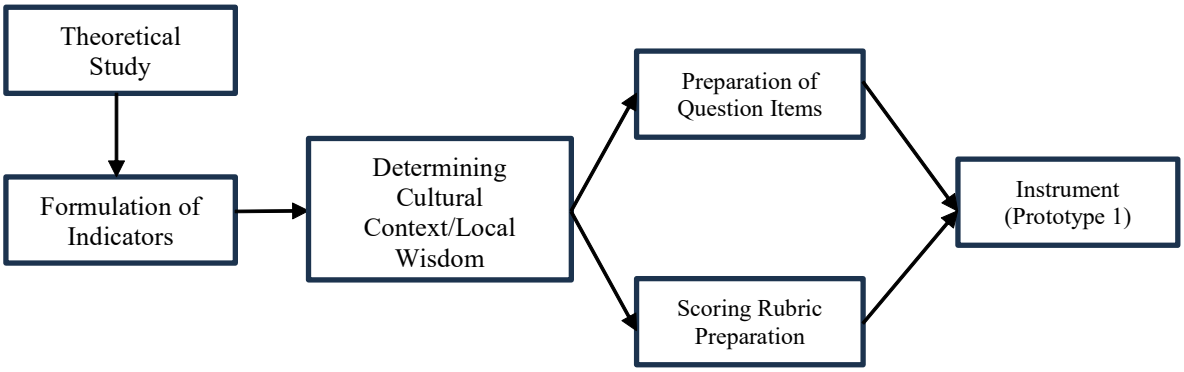

Figure 4. Instrument arrangement diagram

Table 9. Characteristics of numeracy literacy and problem-solving ability instruments

Instrument Type	Question Format	Number of Grains	Minimum Score	Maximum Score
Numeracy Literacy	Complex Multiple Choice	11	0	4
Problem-solving	Description	6	0	4

Table 10. Numeracy literacy ability scoring rubric

Score	Description
0	No answer or 0 correct answers
1	1 correct answer
2	2 correct answers
3	3 correct answers
4	All answers are correct

Table 11. Problem-Solving scoring rubric

Score	Description
0	No answer or a completely irrelevant answer.
1	Shows one step of Polya or an answer in the form of a guess without mathematical reasoning.
2	Shows two Polya steps, but the answer is incomplete or does not produce a correct solution.
3	Shows Polya's three steps and almost correct answer, but there is no checking or reflection yet.
4	Demonstrates Polya's four steps completely and correctly; logical, systematic, and re-checked answers.

Expert Validation (Content Validity)

The content validity stage involves three experts: those in learning evaluation, mathematics education, and culture/local wisdom (traditional gods). The assessment is carried out using a Likert scale of 1-5 (1 = very unsuitable, 2 = unsuitable, 3 = quite suitable, 4 = suitable, and 5 = very suitable) (Avinç & Doðan, 2024; Ramadhani et al., 2025). Aspects that are of concern to the validators are: 1) content construction, 2) language construction, 3) presentation construction (appearance), and 4) integration of culture/local wisdom. This validity analysis focuses not only on overall feasibility but also on content accuracy and the consistency of assessments across validators. Therefore, the validity of the tool is analyzed through three approaches, namely (1) the average score of the validators to determine the general level of validity (Table 12), (2) Aiken's V coefficient to measure the content validity of each indicator (Table 13), and (3) percentage agreement to determine the level of agreement between validators (Table 14).

Table 12. Instruments validation analysis results based on average validator score

Instruments	Expert I	Expert II	Expert III	Averages	Categories
Numeracy Literacy Test	4.83	4.92	4.83	4.86	Very Valid
Problem-Solving Test	5.00	5.00	5.00	5.00	Very Valid
Averages	4.92	4.96	4.92	4.93	Very Valid

Table 13. V aiken analysis results

Instruments	Aspects	Average Scores	n	S1	S2	S3	ΣS	Aiken V-value	Categories
Numeracy Literacy Test	Contents	4.58	3	3.50	3.75	3.50	10.75	0.90	Very Valid
	Presentation	5.00	3	4.00	4.00	4.00	12.00	1.00	Very Valid
	Language	5.00	3	4.00	4.00	4.00	12.00	1.00	Very Valid
Problem-Solving Test	Contents	5.00	3	4.00	4.00	4.00	12.00	1.00	Very Valid
	Presentation	5.00	3	4.00	4.00	4.00	12.00	1.00	Very Valid
	Language	5.00	3	4.00	4.00	4.00	12.00	1.00	Very Valid

Table 14. Percentage agreement analysis results

Instruments	A – B	A + B	$\frac{A-B}{A+B}$	$1-\frac{A-B}{A+B}$	$\left(1-\frac{A-B}{A+B}\right)\times 100\%$	Categories
Numeracy Literacy Test	0.08	9.75	0.01	0.99	99%	Very High Agreement
Tes Pemecahan Masalah	0.00	10.00	0.00	1.00	100%	Very High Agreement
Average					99.50%	Very High Agreement

The customary council itself paid special attention to the integration of culture/local wisdom by assessing the elements used in the instrument, for example, related to the colors of the Donggala woven fabric, the structure of the Souraja house, and others. This input enriched the validation perspective, ensuring that the instrument was not only academically valid but also authentic. Based on validation by cultural experts, the instrument was deemed highly valid in its integration of culture/local wisdom, with an average score of 4.75. This finding indicates that the cultural context integrated in the instrument was conceptually appropriate, authentic, and relevant.

Overall, the results of expert validation show that the numeracy literacy and problem-solving skills instruments developed are declared very valid and suitable for use in the pilot study (small group) (Avinç & Doðan, 2024; Olkun et al., 2016).

Small Group

The small-group instrument trial phase was conducted after the instrument was deemed valid by expert assessment. The trial was conducted to obtain initial empirical data to evaluate the instrument's psychometric quality, particularly with respect to item characteristics and overall measurement quality, before the instrument was used on a larger scale. The instrument was trialed on 71 high school students from three educational units: SMAN 3 Balaesang (Donggala Regency), SMAN 8 Palu (Palu City), and SMAN 8 Sigi (Sigi Regency).

The pilot test was conducted by administering two instruments: a numeracy-literacy skill instrument and a problem-solving skill instrument. All student responses were then coded and formatted for analysis using Rasch modeling. This small-group phase was not intended to test differences between schools, but rather to identify student response patterns, detect potential item weaknesses, and ensure the instrument's initial stability as a measurement tool.

Psychometric Quality Analysis of Instruments using Rasch Modeling

The psychometric quality analysis of the numeracy literacy and problem-solving ability instrument for high school students was conducted

using Rasch modeling in Winsteps version 5.10.1.0. The analysis focused on five main aspects, namely: 1) item fit, 2) distribution of item difficulty levels and respondent abilities, 3) reliability and separation index, 4) instrument unidimensionality, and 5) potential item bias through Differential Item Functioning (DIF) analysis. As an initial stage in the Rasch Model, the Wright Map is used to map the distribution of respondent abilities and item difficulty levels on the same measurement scale (Avinç & Doðan, 2024; Syifa et al., 2024).

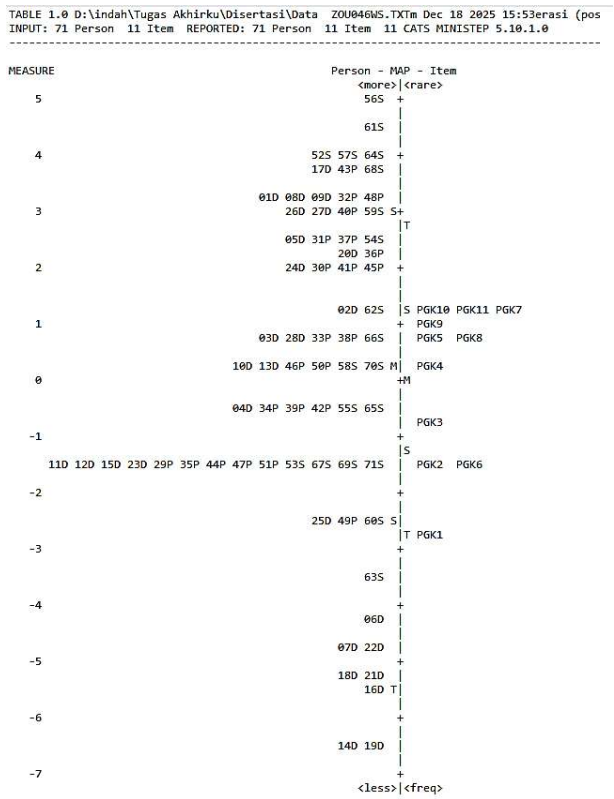

Figure 5. Wright map analysis of numeracy literacy ability instrument

Based on a Rasch analysis of 71 respondents and 11 numeracy literacy items, the Wright Person-Item Map indicates that students' ability levels range from -6.58 to 4.95 logits. Meanwhile, the item difficulty levels are concentrated in a narrower range, from -2.67 to 1.36. This difference in range indicates that the instrument is not yet fully capable of capturing the full range of student abilities, both in the very high- and very low-ability groups, due to the limited number of items at very high and very low difficulty levels. Nevertheless, most students fall within the ability range that overlaps with the item difficulty distribution, so the instrument still provides meaningful measurement information for the dominant ability group in the sample. This finding confirms that the instrument is suitable for use in the initial development stage, but requires the addition of more extreme items to improve targeting quality and measurement precision (Avinç & Doðan, 2024).

Based on Figure 4, the item difficulty level ranges from -2.67 to 1.36 logits, with PGK7, PGK10, and PGK11 as the most difficult items, and PGK1, PGK2, and PGK6 as the easiest items. This distribution indicates that the instrument

TABLE 13.1 D:\indah\Tugas Akhirku\Disertasi\Data ZOU046WS.TXT Dec 18 2025 15:53merasi (pos
INPUT: 71 Person 11 Item REPORTED: 71 Person 11 Item 11 CATS MINISTEP 5.10.1.0

Person: REAL SEP.: 3.25 REL.: .91 ... Item: REAL SEP.: 4.29 REL.: .95

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item	G
7	156	71	1.36	.29	1.34	1.68	1.32	1.04	.65	.73	69.0	78.2	PGK7	4
10	157	71	1.28	.29	.89	-.53	.70	-1.00	.80	.73	81.7	78.2	PGK10	4
11	158	71	1.19	.29	.67	-1.90	.55	-1.66	.82	.73	83.1	78.2	PGK11	4
9	160	71	.91	.31	.92	-.35	.65	-.93	.71	.68	76.1	80.1	PGK9	5
5	162	71	.86	.29	1.20	1.01	1.10	.42	.74	.73	78.9	78.0	PGK5	4
8	164	71	.69	.29	1.35	1.69	1.10	.44	.71	.74	71.8	77.9	PGK8	4
4	171	71	.14	.28	1.14	.74	.89	-.36	.68	.74	71.8	77.1	PGK4	4
3	185	71	-.86	.26	.74	-1.68	.63	-1.97	.79	.75	73.2	71.8	PGK3	4
6	157	71	-1.39	.32	.92	-.42	.84	-.42	.74	.72	81.7	79.8	PGK6	3
2	195	71	-1.52	.25	.98	-.10	.89	-.49	.76	.76	70.4	69.7	PGK2	4
1	213	71	-2.67	.26	1.13	.84	1.25	1.12	.72	.79	71.8	70.3	PGK1	4
MEAN	170.7	71.0	.00	.28	1.03	.09	.90	-.35			75.4	76.3		
P.SD	18.0	.0	1.32	.02	.22	1.16	.25	.97			4.9	3.6		

Figure 6. Item analysis of the numeracy literacy ability instrument

covers an adequate variety of cognitive levels. The fit between the item difficulty level and the distribution of student abilities on the Wright map indicates good instrument targeting, indicating that, empirically, the instrument meets construct validity under the Rasch model.

TABLE 10.1 D:\indah\Tugas Akhirku\Disertasi\Data ZOU046WS.TXT Dec 18 2025 15:53merasi (pos
INPUT: 71 Person 11 Item REPORTED: 71 Person 11 Item 11 CATS MINISTEP 5.10.1.0

Person: REAL SEP.: 3.25 REL.: .91 ... Item: REAL SEP.: 4.29 REL.: .95

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item	G
8	164	71	.69	.29	1.35	1.69	1.10	.44	A .71	.74	71.8	77.9	PGK8	4
7	156	71	1.36	.29	1.34	1.68	1.32	1.04	B .65	.73	69.0	78.2	PGK7	4
1	213	71	-2.67	.26	1.13	.84	1.25	1.12	C .72	.79	71.8	70.3	PGK1	4
5	162	71	.86	.29	1.20	1.01	1.10	.42	D .74	.73	78.9	78.0	PGK5	4
4	171	71	.14	.28	1.14	.74	.89	-.36	E .68	.74	71.8	77.1	PGK4	4
2	195	71	-1.52	.25	.98	-.10	.89	-.49	F .76	.76	70.4	69.7	PGK2	4
6	157	71	-1.39	.32	.92	-.42	.84	-.42	e .74	.72	81.7	79.8	PGK6	3
9	160	71	.91	.31	.92	-.35	.65	-.93	d .71	.68	76.1	80.1	PGK9	5
10	157	71	1.28	.29	.89	-.53	.70	-1.00	c .80	.73	81.7	78.2	PGK10	4
3	185	71	-.86	.26	.74	-1.68	.63	-1.97	b .79	.75	73.2	71.8	PGK3	4
11	158	71	1.19	.29	.67	-1.90	.55	-1.66	a .82	.73	83.1	78.2	PGK11	4
MEAN	170.7	71.0	.00	.28	1.03	.09	.90	-.35			75.4	76.3		
P.SD	18.0	.0	1.32	.02	.22	1.16	.25	.97			4.9	3.6		

Figure 7. Item fit order analysis of numeracy literacy ability instrument

The results of the analysis show that all items meet the suitability criteria of the Rasch model, with the Infit and Outfit MNSQ values in the range of 0.5–1.5 and a positive point measure correlation (Ariffin et al., 2010). These findings indicate that the instrument is unidimensional and that each item consistently measures the construct of numeracy literacy. Furthermore, a Differential Item Functioning (DIF) analysis was conducted to assess the equivalence of item functioning across respondent groups by region.

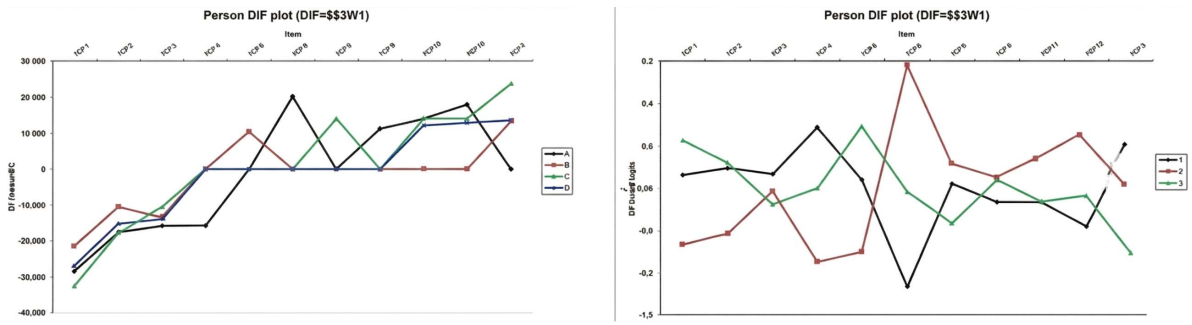

Figure 8. Person DIF plot analysis

Based on the Person DIF Plot in Figure 6, several items showed more striking functional differences in the Sigi Regency group compared to the Palu and Donggala groups. To follow up on these findings, a content review was conducted of the items that indicated DIF. The review results indicated that potential bias was not attributable to the numeracy literacy or problem-solving constructs measured but was influenced by non-construct factors, such as familiarity with the item stimuli's cultural context, language complexity, and differences in students' contextual experiences across regions. This condition aligns with the profile of numeracy literacy achievement in Sigi Regency, which remains relatively low compared with the Palu and Donggala regions, according to the Education Report Card. Consequently, students in Sigi may face additional cognitive load at the conceptual understanding stage of the questions. In line with psychometric literature, DIF findings at this stage are positioned as diagnostic information for instrument improvement through revision of the context and language of the items, rather than as a fatal weakness of the instrument. Therefore, the DIF results confirm that the instrument is feasible at the initial development stage (pilot study), but still requires revision and revalidation before being used across regions (Avinç & Doğan, 2024; Boone & Staver, 2020; Debelak et al., 2022; Winckel et al., 2022)

TABLE 3.1 D:\indah\Tugas Akhirku\Disertasi\Data ZOU046WS.TXTm Dec 18 2025 15:53erasi (pos
INPUT: 71 Person 11 Item REPORTED: 71 Person 11 Item 11 CATS MINISTEP 5.10.1.0

SUMMARY OF 71 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	26.5	11.0	.21	.76	.95	.01	.90	.01
SEM	.6	.0	.34	.02	.06	.12	.07	.12
P.SD	5.3	.0	2.83	.16	.49	1.02	.60	.99
S.SD	5.4	.0	2.85	.16	.50	1.03	.60	.99
MAX.	37.0	11.0	4.95	1.04	2.32	2.56	2.44	2.52
MIN.	15.0	11.0	-6.58	.59	.13	-1.65	.08	-1.72
REAL RMSE	.83 TRUE SD	2.71	SEPARATION	3.25	Person RELIABILITY	.91		
MODEL RMSE	.77 TRUE SD	2.72	SEPARATION	3.52	Person RELIABILITY	.93		
S.E. OF Person MEAN = .34								

Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .92 SEM = 1.49
STANDARDIZED (50 ITEM) RELIABILITY = .98

Figure 9. Reliability analysis of person numeracy literacy ability

SUMMARY OF 11 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	170.7	71.0	.00	.28	1.03	.09	.90	-.35
SEM	5.7	.0	.42	.01	.07	.37	.08	.31
P.SD	18.0	.0	1.32	.02	.22	1.16	.25	.97
S.SD	18.9	.0	1.38	.02	.23	1.22	.26	1.02
MAX.	213.0	71.0	1.36	.32	1.35	1.69	1.32	1.12
MIN.	156.0	71.0	-2.67	.25	.67	-1.90	.55	-1.97
REAL RMSE	.30 TRUE SD	1.28	SEPARATION	4.29	Item RELIABILITY	.95		
MODEL RMSE	.28 TRUE SD	1.29	SEPARATION	4.52	Item RELIABILITY	.95		
S.E. OF Item MEAN = .42								

Item RAW SCORE-TO-MEASURE CORRELATION = -.83 (approximate due to multiple item groupings)
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000

Figure 10. Reliability analysis of numeracy literacy ability items

Based on summary statistics, the instrument demonstrated excellent reliability, with a person reliability of 0.91 and an item reliability of 0.95. The person separation (3.25) and item separation (4.29) values demonstrate the instrument's ability to clearly differentiate student ability levels and

item difficulty levels. A Cronbach's Alpha (KR-20) of 0.92 indicates very high internal consistency, making the instrument suitable for research and learning evaluation (Raof et al., 2021). Furthermore, the Rasch modeling analysis focused on students' mathematical problem-solving abilities to obtain an overview of item characteristics and the distribution of respondents' abilities on the construct.

Based on the Wright map, the distribution of respondents' abilities and item difficulty levels is relatively balanced along a logit continuum.

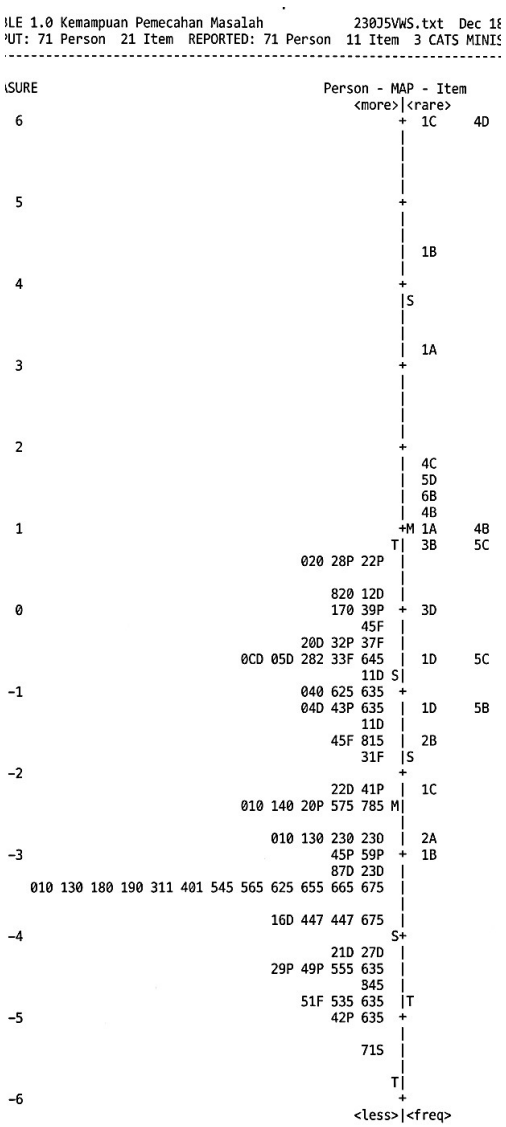

Figure 11. Wright map analysis problem-solving ability instrument

Respondents' abilities range from approximately "6 to +6 logits, with most falling between e1 and +3 logits, indicating heterogeneous problem-solving abilities and good instrument sensitivity in differentiating individual differences.

Based on the item-measure analysis in Figure 10, the difficulty level of the problem-solving instrument items ranges from -2.93 to 6.92 logits. However, the distribution of item difficulty is uneven, as indicated by a significant

TABLE 13.1 Kemampuan Pemecahan Masalah ZOU187WS.TXT Dec 18 2025 23:05
INPUT: 71 Person 21 Item REPORTED: 71 Person 21 Item 3 CATS MINISTEP 5.10.1.0

Person: REAL SEP.: 2.27 REL.: .84 ... Item: REAL SEP.: 3.62 REL.: .93

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
14	47	47	6.92	1.83	MAXIMUM MEASURE				.00	.00	100.0	100.0	4d
9	35	35	6.88	1.83	MAXIMUM MEASURE				.00	.00	100.0	100.0	3c
21	34	34	6.86	1.84	MAXIMUM MEASURE				.00	.00	100.0	100.0	6c
19	52	45	3.23	.45	.94	-.17	.60	-.34	.45	.40	86.7	84.7	6a
13	89	70	1.85	.33	1.97	4.31	2.83	3.28	.13	.56	65.7	80.7	4c
18	78	59	1.67	.34	1.51	2.46	1.46	1.19	.34	.57	62.7	78.9	5d
20	76	54	1.40	.33	.66	-2.08	.51	-1.78	.73	.57	83.3	76.5	6b
12	97	71	1.16	.31	1.68	3.21	1.88	2.48	.26	.60	60.6	79.6	4b
7	99	71	.97	.30	.40	-4.36	.32	-3.40	.86	.61	97.2	79.2	3a
11	99	71	.97	.30	.40	-4.36	.32	-3.40	.86	.61	97.2	79.2	4a
1	100	71	.88	.30	.39	-4.48	.30	-3.65	.87	.61	97.2	79.0	1a
15	100	71	.88	.30	.63	-2.38	.46	-2.47	.77	.61	83.1	79.0	5a
10	110	71	.04	.28	1.23	1.38	1.35	1.53	.75	.60	71.8	75.7	3d
17	86	47	-.58	.35	.88	-.49	.96	-.04	.20	.46	78.7	75.6	5c
8	117	70	-.59	.28	.51	-3.82	.41	-3.74	.72	.57	81.4	72.8	3b
2	125	71	-1.11	.27	.60	-2.93	.50	-3.03	.58	.55	83.1	73.6	1b
16	93	48	-1.27	.35	.51	-2.47	.55	-1.82	.11	.45	95.8	78.8	5b
6	131	71	-1.56	.27	1.84	4.29	2.11	4.38	.36	.54	52.1	74.0	2b
3	140	71	-2.24	.28	1.64	3.26	1.74	3.04	.57	.53	60.6	74.7	1c
5	147	71	-2.78	.28	1.03	.22	1.05	.32	.60	.54	76.1	75.7	2a
4	149	71	-2.93	.28	.76	-1.42	.70	-1.47	.40	.55	81.7	76.0	1d
MEAN	95.4	61.4	.98	.53	.98	-.55	1.00	-.49			78.6	77.4	
P.SD	33.0	13.0	2.87	.53	.52	2.98	.72	2.57			13.5	2.9	

Figure 12. Item analysis of problem-solving ability instrument measure

logit jump between items with moderate difficulty and those with very high difficulty (logits> +3). This gap indicates that measurement of problem-solving ability has not been conducted consistently within the medium-high ability range.

Substantively, the logit jump indicates that very difficult items likely demand different cognitive processes, such as multi-step modeling, integration of multiple representations, and high-level abstract reasoning, and thus do not fully align with items of moderate difficulty. Therefore, further development is focused on constructing items with a medium-high difficulty level (0 to +3 logit) through gradual increases in complexity, without jumping to high abstraction demands, in order to improve measurement continuity and instrument targeting (Boone & Staver, 2020; Debelak et al., 2022; Raof et al., 2021; Winckel et al., 2022).

The results of the item fit analysis indicate that most items have Infit and Outfit MNSQ values that are close to the Rasch model

TABLE 10.1 Kemampuan Pemecahan Masalah ZOU187WS.TXT Dec 18 2025 23:05
INPUT: 71 Person 21 Item REPORTED: 71 Person 21 Item 3 CATS MINISTEP 5.10.1.0

Person: REAL SEP.: 2.27 REL.: .84 ... Item: REAL SEP.: 3.62 REL.: .93

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
13	89	70	1.85	.33	1.97	4.31	2.83	3.28	A .13	.56	65.7	80.7	4c
6	131	71	-1.56	.27	1.84	4.29	2.11	4.38	B .36	.54	52.1	74.0	2b
12	97	71	1.16	.31	1.68	3.21	1.88	2.48	C .26	.60	60.6	79.6	4b
3	140	71	-2.24	.28	1.64	3.26	1.74	3.04	D .57	.53	60.6	74.7	1c
18	78	59	1.67	.34	1.51	2.46	1.46	1.19	E .34	.57	62.7	78.9	5d
10	110	71	.04	.28	1.23	1.38	1.35	1.53	F .75	.60	71.8	75.7	3d
5	147	71	-2.78	.28	1.03	.22	1.05	.32	G .60	.54	76.1	75.7	2a
17	86	47	-.58	.35	.88	-.49	.96	-.04	H .20	.46	78.7	75.6	5c
19	52	45	3.23	.45	.94	-.17	.60	-.34	I .45	.40	86.7	84.7	6a
4	149	71	-2.93	.28	.76	-1.42	.70	-1.47	i .40	.55	81.7	76.0	1d
20	76	54	1.40	.33	.66	-2.08	.51	-1.78	h .73	.57	83.3	76.5	6b
15	100	71	.88	.30	.63	-2.38	.46	-2.47	g .77	.61	83.1	79.0	5a
2	125	71	-1.11	.27	.60	-2.93	.50	-3.03	f .58	.55	83.1	73.6	1b
16	93	48	-1.27	.35	.51	-2.47	.55	-1.82	e .11	.45	95.8	78.8	5b
8	117	70	-.59	.28	.51	-3.82	.41	-3.74	d .72	.57	81.4	72.8	3b
7	99	71	.97	.30	.40	-4.36	.32	-3.40	c .86	.61	97.2	79.2	3a
11	99	71	.97	.30	.40	-4.36	.32	-3.40	b .86	.61	97.2	79.2	4a
1	100	71	.88	.30	.39	-4.48	.30	-3.65	a .87	.61	97.2	79.0	1a
MEAN	95.4	61.4	.98	.53	.98	-.55	1.00	-.49			78.6	77.4	
P.SD	33.0	13.0	2.87	.53	.52	2.98	.72	2.57			13.5	2.9	

Figure 13. Item fit order analysis of problem-solving ability instrument

acceptance criteria (0.5–1.5) (Ariffin et al., 2010). Although there are several items with MNSQ values outside the ideal range, the deviation is not extreme and is still tolerable. Overall, the items demonstrated adequate model fit and supported construct validity; therefore, further DIF analysis was conducted to assess the equivalence of item functions across regions.

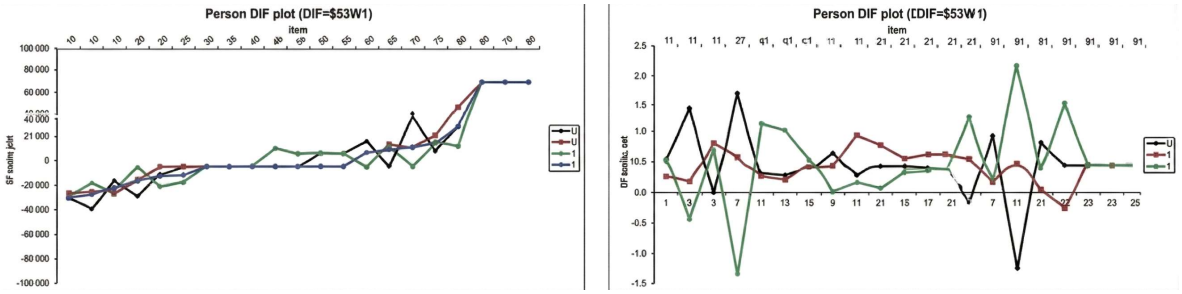

Figure 14. Person DIF plot analysis

The DIF analysis results in Figure 12 indicate differences in item functioning across regions, with respondents from Palu and Donggala having a higher probability of success than those from Sigi Regency on several items. This difference is likely related to variations in Education Report Card achievement across regions, which reflect differences in learning contexts and students' levels of exposure to problem-solving experiences. This pattern indicates the potential for systematic interregional bias, whereby differences in responses do not entirely reflect differences in latent ability but are also influenced by non-construct factors. In line with the principles of Rasch-based assessment, items with strong indications of DIF require follow-up through content review to identify sources of bias, whether related to the cultural context of the stimulus, language complexity, or the cognitive demands of students' learning experiences across regions. Therefore, the findings in Figure 12 serve as an empirical basis for instrument revision and refinement in the next stage of development, to ensure fairness and equity of measurement before the instrument is implemented across regions (Avinç & Doğan, 2024; Boone & Staver, 2020; Debelak et al., 2022; Winckel et al., 2022).

```
TABLE 3.1 Kemampuan Pemecahan Masalah          ZOU187WS.TXT  Dec 18 2025 23:05
INPUT: 71 Person  21 Item  REPORTED: 71 Person  21 Item  3 CATS MINISTEP 5.10.1.0
-------------------------------------------------------------------------------

    SUMMARY OF 71 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE     S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      28.2      18.2       -2.43      .62       .98   -.07   1.03    .10 |
|  SEM        .9        .3         .20      .01       .04    .12    .07    .12 |
| P.SD       7.8       2.7        1.64      .08       .35   1.03    .56   1.01 |
| S.SD       7.9       2.8        1.66      .08       .35   1.04    .56   1.02 |
| MAX.      41.0      21.0         .58      .90      1.94   2.30   2.54   2.12 |
| MIN.      14.0      12.0       -5.50      .54       .41  -2.28    .22  -2.14 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .66 TRUE SD    1.51  SEPARATION  2.27  Person RELIABILITY  .84 |
|MODEL RMSE    .63 TRUE SD    1.52  SEPARATION  2.43  Person RELIABILITY  .86 |
| S.E. OF Person MEAN = .20                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION =  .98 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .98  SEM = 1.20 (approximate due to missing data)
STANDARDIZED (50 ITEM) RELIABILITY = .93
```

Figure 15. Reliability analysis of a person's problem-solving ability

SUMMARY OF 18 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	104.9	65.2	.00	.31	.98	-.55	1.00	-.49
SEM	6.1	2.3	.41	.01	.13	.72	.17	.62
P.SD	25.3	9.5	1.68	.04	.52	2.98	.72	2.57
S.SD	26.0	9.7	1.73	.04	.54	3.07	.74	2.65
MAX.	149.0	71.0	3.23	.45	1.97	4.31	2.83	4.38
MIN.	52.0	45.0	-2.93	.27	.39	-4.48	.30	-3.74

REAL RMSE	.34 TRUE SD	1.64	SEPARATION	4.77	Item	RELIABILITY	.96
MODEL RMSE	.31 TRUE SD	1.65	SEPARATION	5.24	Item	RELIABILITY	.96
S.E. OF Item MEAN = .41							

MINIMUM EXTREME SCORE: 3 Item 14.3%

Figure 16. Reliability analysis of problem-solving ability items

Summary statistics indicate that the problem-solving instrument has person reliability of 0.84–0.86 and item reliability of 0.96, indicating measurement consistency and stability of the item difficulty hierarchy. Person separation values (2.27–2.43) and item separation values (4.77–5.24) demonstrate the instrument's ability to differentiate respondents' ability levels and to group items by difficulty. However, the Cronbach's Alpha (KR-20) value of 0.98 confirms very high internal consistency and should be interpreted with caution. In a psychometric context, very high reliability may indicate functional similarity among items or item redundancy, potentially limiting the breadth of construct representation (Boone & Staver, 2020; Debelak et al., 2022; Winckel et al., 2022).

To follow up on these findings, a Principal Component Analysis of Residuals (PCAR) analysis was conducted on the Rasch modeling output to examine the unidimensionality assumption more deeply. The PCAR results showed that the variance explained by the primary dimensions reached 50.4%, with essential unidimensionality at 65.7%, indicating the dominance of the primary latent construct. However, the eigenvalue for the first contrast (3.92) indicated the potential for secondary dimensions or functional similarities between items. This finding served as diagnostic evidence during the pilot study phase, underscoring the need to reduce redundancy and broaden the range of cognitive demands in subsequent instrument development. Nevertheless, the instrument was deemed suitable for further development (Raof et al., 2021).

The measuring capability of the numeracy literacy and problem-solving instruments is assessed through analyses of content and empirical validity. Content validity is achieved through expert assessment, which translates qualitative considerations into objective, measurable quantitative evidence. The reliability of this process depends heavily on the competence and consistency of the experts in assessing construct suitability, indicator clarity, and the accuracy of cultural/local wisdom integration, thereby minimizing potential bias (Avinç & Doðan, 2024; Reffiane et al., 2021). Furthermore, empirical validity was analyzed using Rasch modeling. The results of Rasch modeling indicate that the numeracy literacy and problem-solving ability instruments have adequate psychometric properties for early-stage development (Avinç & Doðan, 2024; Nisa et al., 2024; Ramadhani et al., 2025; Raof et al., 2021). Most items meet the item fit criteria and support construct validity, while person and item reliability fall within the good to very good range (Avinç & Doðan, 2024; Raof et al., 2021). However, these

findings should be interpreted comprehensively, with the PCAR results serving as an additional diagnostic test of the unidimensionality assumption (Boone & Staver, 2020; Debelak et al., 2022). The PCAR results indicate that the variance explained by the primary dimensions exceeds the recommended minimum. However, the eigenvalues in the first contrast indicate the potential for secondary dimensions or functional similarities between items, particularly in problem-solving. This finding does not necessarily negate the validity of the construct. However, it suggests that some items still exhibit similar residual patterns, indicating that the breadth of construct representation is not yet fully optimal at this stage (Boone & Staver, 2020; Debelak et al., 2022; Raof et al., 2021; Winckel et al., 2022).

Consistent with the PCAR findings, Wright's map indicates that the distribution of item difficulty levels does not fully align with the distribution of respondents' abilities, particularly at the very high and very low levels. Furthermore, the problem-solving instrument still exhibits significant differences in difficulty between items, so that ability measurement does not occur continuously across the entire ability range (Avinç & Doðan, 2024; Syifa et al., 2024). The cross-regional DIF findings also suggest that some items may be influenced by local context, such that item functioning is not completely equivalent across groups (Avinç & Doðan, 2024). This condition strengthens the interpretation of the PCAR results that, in addition to the structural aspect of unidimensionality, the instrument also faces challenges in the aspects of instrument targeting, continuity of difficulty levels, and fairness across measurement contexts (Boone & Staver, 2020; Debelak et al., 2022; Winckel et al., 2022).

The main limitations of this study lie in instrument targeting, the continuity of difficulty levels, and fairness across measurement contexts. The very high internal reliability values of the problem-solving instruments also need to be interpreted critically because they potentially reflect functional similarities between items and limitations in construct breadth (Avinç & Doðan, 2024; Syifa et al., 2024). Therefore, further research is recommended to develop extreme items, refine the difficulty gradient, and revise the instrument based on DIF findings to improve its fairness and precision. Further validation with a broader, more diverse sample, as well as testing across different cultural contexts, is needed to ensure that the instrument can be used more generally and sustainably to measure numeracy literacy and problem-solving.

■ CONCLUSION

This study shows that the developed numeracy literacy and problem-solving ability instrument has generally adequate psychometric quality, as indicated by the fit of most items to the model, high person and item reliability, and the instrument's ability to differentiate respondents' ability levels. However, the findings also indicate limitations in the instrument's targeting, particularly a disparity in item difficulty at the extremes of the ability range and indications of potential inter-regional bias in some items. Therefore, claims of instrument suitability should be understood proportionally as valid and reliable within the context of the research being tested. However, they still require refinement to ensure more comprehensive coverage of constructs and measurement fairness across a wider population.

Based on these findings, further research is recommended to develop additional items with more varied difficulty levels, particularly items capable of measuring students with very high and very low ability, to improve the instrument's targeting quality. Furthermore, DIF analysis should be followed by revising or replacing items identified as biased to ensure the instrument can be used fairly across regional contexts. Further testing with a larger, more heterogeneous sample is also recommended to strengthen the

generalizability of the results and to ensure that the high reliability observed is not due to item redundancy but rather reflects the breadth and depth of the construct being measured.

DISCLOSURE OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the writing of this manuscript, the author employed ChatGPT to assist with language refinement/proofreading. The author has reviewed and edited the content generated by this tool and assumes full responsibility for the content of the published article.

REFERENCES

Adams, E. K., Hancock, K. J., & Taylor, C. L. (2020). Student achievement against national minimum standards for reading and numeracy in years 3, 5, 7, and 9: A regression discontinuity analysis. *Australian Journal of Social Issues*, *55*(3), 275–301. https://doi.org/10.1002/ajs4.124

Amalina, I. K., & Vidákovich, T. (2022). An integrated STEM-based mathematical problem-solving test: Developing and reporting psychometric evidence. *Journal on Mathematics Education*, *13*(4), 587–604. https://doi.org/http://doi.org/10.22342/jme.v13i4.pp587-604

Amalina, I. K., & Vidákovich, T. (2023). Assessment of domain-specific prior knowledge/ : A development and validation of mathematical problem-solving test. *International Journal of Evaluation and Research in Education*, *12*(1), 468–476. https://doi.org/10.11591/ijere.v12i1.23831

Andrian, D., Nofriyandi, Wahyuni, R., Loska, F., Nurhalimah, S., Mz, Z. A., Risnawati, Maisaroh, S., Dacara, E., & Hidayat, R. (2025). Numeracy skills students' profile of Riau Province: content and construct instrument-based Riau culture and evaluation results. *International Journal of Evaluation and Research in Education*, *14*(6), 4638–4654. https://doi.org/10.11591/ijere.v14i6.33426

Anriana, R., Witri, G., Putra, Z. H., Fendrik, M., Dahnilsyah, & Aljarrah, A. (2023). Ethnomathematics study in measurement of Bengkalis Malay community as mathematics resources for elementary school. *Ethnography and Education*, *18*(3), 299–322. https://doi.org/10.1080/17457823.2023.2232500

Ariffin, S. R., Omar, B., Isa, A., & Sharif, S. (2010). Validity and reliability multiple intelligent item using rasch measurement model. *Procedia: Social and Behavioral Sciences*, *9*, 729–733. https://doi.org/10.1016/j.sbspro.2010.12.225

Asy'ari, M., Ikhsan, M., & Muhali, M. (2018). Validitas instrumen karakterisasi kemampuan metakognisi mahasiswa calon guru fisika. *Prisma Sains/ : Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram*, *6*(1), 18. https://doi.org/10.33394/j-ps.v6i1.955

Avinç, E., & Doðan, F. (2024). Digital literacy scale/ : Validity and reliability study with the rasch model. In *Education and Information Technologies* (Vol. 29, Issue 17). Springer US. https://doi.org/10.1007/s10639-024-12662-7

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Springer Nature Switzerland.

Charoentham, M., Kantathanawat, T., Pimdee, P., & Teerasoonthontai, K. (2025). Assessing creative problem-solving competencies in primary school educators/ : A confirmatory factor analysis approach. *Edelweiss Applied Science and Technology*, *9*(1), 262–277. https://

doi.org/10.55214/25768484.v9i1.4111

Chen, Y., Guo, C. J., Lim, K. M., Mun, K. J., Otsuji, H., Park, Y. S., Sorrell, D., & Winnie So, W. M. (2020). The influence of school entry skills in literacy and numeracy on the science achievement of fourth grade students and schools in Asian regions. *Eurasia Journal of Mathematics, Science and Technology Education*, *16*(9). https://doi.org/10.29333/EJMSTE/8385

Debelak, R., Strobl, C., & Zeigenfuse, M. D. (2022). *An introduction to the rasch model with examples in r*. Boca Raton: CRC Press.

Directorate of Senior High School. (2021). *Panduan penguatan literasi dan numerasi* [Guide to strengthening literacy and numeracy]. Ministry of Education and Culture.

Evans, C. M., & Taylor, C. S. (2025). *Culturally responsive assessment in classrooms and large-scale contexts*. New York: Routledge.

Gusmanida, G., Sujati, H., & Herwin. (2024). Testing the construct validity and reliability of the student learning motivation scale using confirmatory factor analysis (cfa). *Jurnal Penelitian Pendidikan IPA*, *10*(7), 4227–4234. https://doi.org/https://doi.org/10.29303/jppipa.v10i7.7518

Hardiyanti, T. A., Syaf, A. H., & Widiastuti, T. (2022). Pengembangan modul berbasis etnomatematika pada materi barisan dan deret. *Seminar Nasional Pendidikan, FKIP UNMA*, 285–300.

Hellstrand, H., Korhonen, J., Räsänen, P., Linnanmäki, K., & Aunio, P. (2020). Reliability and validity evidence of the early numeracy test for identifying children at risk for mathematical learning difficulties. *International Journal of Educational Research*, *102*, 101580. https://doi.org/https://doi.org/10.1016/j.ijer.2020.101580

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Husain, H., & Aziz, H. (2022). Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were used to assess the construct validity and reliability of historical thinking skills, TPACK, and the application of historical thinking skills. *International Journal of Education, Psychology and Counselling*, *7*(46), 608–623. https://doi.org/10.35631/IJEPC.746046

Jackson, C. J. (2022). The utility of NAPLAN data: issues of access, use and expertise for teaching and learning. *Australian Journal of Language and Literacy*, *45*(2), 141–157. https://doi.org/10.1007/s44020-022-00009-z

Kline, R. B. (2016). *Principles and practice of Structural Equation Modeling*. New York: The Guilford Press.

Ladson-billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465–491.

Lam, T. T., Seng, Q. K., Hoong, L. Y., Jaguthsing, D., & Guan, T. E. (2011). Making mathematics practical: An approach to problem solving. *Making Mathematics Practical: An Approach to Problem Solving*, 1–156. https://doi.org/10.1142/8171

Lin, Y., Yu, Y., Zeng, J., Zhao, X., & Wan, C. (2020). Comparing the reliability and validity of the SF-36 and SF-12 in measuring quality of life among adolescents in China: a large sample cross-sectional study. *Health and Quality of Life*

Outcomes, *18*(1), 1–14. https://doi.org/10.1186/s12955-020-01605-8

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science*, *26*(1–2), 49–63. https://doi.org/10.1023/a:1003088013286

Ministry of Education, Culture, Research, and Technology. (2023). *Framework asesmen kompetensi minimum (AKM)* [Minimum competency assessment (AKM) framework]. Center for Educational Assessment, Agency for Standardization, Curriculum, and Assessment in Education.

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia: Social and Behavioral SciencesSocial and Behavioral Sciences*, *204*, 164–171. https://doi.org/10.1016/j.sbspro.2015.08.129

Nahdi, D. S., Jatisunda, M. G., Cahyaningsih, U., & Suciawati, V. (2020). Pre-service teacher's ability in solving mathematics problem viewed from numeracy literacy skills. *Elementary Education Online*, *19*(4), 1902–1910. https://doi.org/10.17051/ilkonline.2020.762541

Nisa, K., Suprapto, N., Amiruddin, M. Z., Sari, E. P. D. N., & Athiah, B. D. (2024). Ethnoscience-quizizz test to measure problem-solving skills/ : a Rasch analysis. *International Journal of Evaluation and Research in Education*, *13*(6), 4247–4255. https://doi.org/10.11591/ijere.v13i6.28075

OECD. (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD survey of adults skills*. OECD Publishing. https://doi.org/10.1787/9789264128859-en

OECD. (2023). *Pisa 2022 results (volume I): The state of learning and equity in education*. PISA, OECD Publishing, Paris. https://doi.org/10.1787/53f23881-en

Olkun, S., Altun, A., ªahin, S. G., & Kaya, G. (2016). Psychometric properties of a screening tool for elementary school students' math learning disorder risk. *International Journal of Learning, Teaching and Educational Research*, *15*(12), 48–66.

Platas, L. M., Ketterlin-Gellar, L., Brombacher, A., & Sitabkhan, Y. (2014). *Early grade mathematics assessment (EGMA) toolkit* (Issue March). RTI International.

Prasetyo, Z. K., Sesa, E., & Lembah, G. (2020). Psychometric and Structural Evaluation of the Physics Metacognition Inventory Instrument. *European Journal of Educational Research*, *9*(1), 215–225.

Polya, G. (1957). How to solve it: a new aspect of mathematical method, second edition. In *Princeton University Press: United States of America* (Vol. 2, p. 253). http://www.jstor.org/stable/3609122?origin=crossref

Pratama, R. A., & Yelken, T. Y. (2024). Effectiveness of ethnomathematics-based learning on students' mathematical literacy: a meta-analysis study. *Discover Education*, *3*(1). https://doi.org/10.1007/s44217-024-00309-1

Rahman, N. A., Masuwai, A., Tajudin, N. M., Tek, O. E., & Adnan, M. (2016). Validation of teaching and learning guiding principles instrument for Malaysian higher learning institutions. *Malaysian Journal of Learning and Instruction*, *13*(2), 125–146. https://doi.org/https://doi.org/10.32890/mjli2016.13.2.5

Rahmawati, D. E., & Trimulyono, G. (2021).

Validitas instrumen penilaian higher order thinking skills (hots) pada materi keanekaragaman hayati. *Berkala Ilmiah Pendidikan Biologi (BioEdu)*, *11*(1), 138–147. https://doi.org/10.26740/bioedu.v11n1.p138-147

Ramadhani, R., Soeharto, S., Arifiyanti, F., Prahmana, R. C. I., Saleh, A., & Lavicza, Z. (2025). Assessing quality and biases in ethnomathematics-based numeracy worksheets: A many-facet Rasch model analysis. *Social Sciences & Humanities Open*, *12*, 101736. https://doi.org/https://doi.org/10.1016/j.ssaho.2025.101736

Raof, S. A., Musta'amal, A. H., Zamzuri, F. K., & Salleh, M. H. (2021). Validity and reliability of students' perceptions on obe approach in Malaysian VC using rasch model. *Journal of Innovation in Educational and Cultural Research*, *2*(2), 44–50. https://doi.org/10.46843/jiecr.v2i2.30

Reffiane, F., Sudarmin, Wiyanto, & Saptono, S. (2021). Developing an instrument to assess students' problem-solving ability on hybrid learning model using ethno-stem approach through quest program. *Pegem Journal of Education and Instruction*, *11*(4), 1–8. https://doi.org/10.47750/pegegog.11.04.01

Rosa, M., D'Ambrosio, U., Orey, D. C., Shirley, L., Alangui, W. V., Palhares, P., & Gavarrete, M. E. (2016). Current and future perspectives of ethnomathematics as a program. In *ICME-13 Topical Surveys*. Switzerland: Springer Nature. http://www.springer.com/series/14352

Sari, C. M., Ridwan, A., & Sastrawijaya, Y. (2024). Learning environment of numeracy: development and validation with confirmatory factor analysis and rasch model. *Pedagogika*, *153*(1), 185–198. https://doi.org/https://doi.org/10.15823/p.2024.153.9 Pedagogika

Schouten, H. (1985). *Statistical measurement of interobserver agreement*. Erasmus Rotterdam.

Setiyani, Waluya, S. B., Sukestiyarno, Y. L., Cahyono, A. N., & Santi, D. P. D. (2024). Assessing numeracy skills on flat shapes and scaffolding forms in junior high school. *International Journal of Evaluation and Research in Education* , *13*(1), 422–432. https://doi.org/10.11591/ijere.v13i1.25186

Suciati, Munadi, S., Sugiman, & Ratna Febriyanti, W. D. (2020). Design and validation of mathematical literacy instruments for assessment for learning in Indonesia. *European Journal of Educational Research*, *9*(2), 865–875. https://doi.org/10.12973/eu-jer.9.2.865

Swarni, A., Herwin, & Sujati. (2024). Testing the construct validity and reliability of the student learning interest scale using confirmatory factor analysis (cfa). *Jurnal Penelitian Pendidikan IPA*, *10*(9), 6322–6330. https://doi.org/https://doi.org/10.29303/jppipa.v10i9.8794

Syifa, A., Azizah, A. N., Fardanti, A. C., Rahayu, F., Indratno, T. K., Sukarelawan, M. I., & Abdullah, N. S. Y. (2024). Problem-solving ability of high school students/ : Preliminary study analysis using Rasch modeling. *Journal of Environment and Sustainability Education*, *2*(2), 135–145. https://doi.org/10.62672/joease.v2i2.37

Umbara, U., & Suryadi, D. (2019). Re-interpretation of mathematical literacy based on the teacher's perspective. *International Journal of Instruction*, *12*(4), 789–806. https://doi.org/10.29333/iji.2019.12450a

Utami, N., Setiawan, A., Hamidah, I., & Koehler, T. (2025). Investigating problem-based

worksheets (PBWs) to improve understanding in logic gates topic/ : stacking and racking analyses in rasch model. *Educational Process International Journal*, *16*, e2025243. https://doi.org/https://doi.org/10.22521/edupij.2025.16.243

Walker, M. E., Olivera-aguilar, M., Lehman, B., Laitusis, C., Guzman-orth, D., & Gholson, M. (2023). *Culturally Responsive Assessment: Provisional Principles* (Issue December). https://doi.org/10.1002/ets2.12374

Winckel, A. van de, Kozlowski, A. J., Johnston, M. V., Weaver, J., Grampurohit, N., Terhorst, L., Juengst, S., Ehrlich-jones, L., Heinemann, A. W., Melvin, J., Sood, P., & Mallinson, T. (2022). Reporting guideline for RULER: Rasch reporting guideline for rehabilitation research: Explanation and elaboration. *Archives of Physical Medicine and Rehabilitation*, *103*(7), 1487–1498. https://doi.org/10.1016/j.apmr.2022.03.019

Zainil, M., Putra, A. P., Sa'ud, U. S., Helsa, Y., Desmaiyanti, Ariani, Y., & Rusdinal. (2024). Development of learning management system model and numerical literacy-based message content. *International Journal of Instruction*, *16*(4), 1041–1060. https://doi.org/https://doi.org/10.29333/iji.2023.16457a

Zebua, V., Rahmi, & Yusri, R. (2020). *Analisis kesalahan siswa dalam menyelesaikan soal barisan dan deret ditinjau dari kemampuan pemahaman konsep matematis* [Analysis of students' errors in solving sequence and series problems viewed from the perspective of their ability to understand mathematical concepts.]. *Jurnal LEMMA*, *6*(2), 122–133. https://doi.org/10.22202/jl.2020.v6i2.4088