



Constructing an instrument to measure and evaluate problem-solving skills in elementary school students: A focus on mathematical word problem-solving

Sepni Yanti^{a,*}, Awaluddin Tjalla^{b,†}, Anan Sutisna^{b,†}, Tri Wahyudi Ramdhan^{c,†}

^a Department of Information Technology, Faculty of Engineering, Mathematics and Natural Sciences, Indraprasta PGRI University, 12530, Indonesia

^b Doctoral Program Department of Educational Research and Evaluation, Jakarta State University, Jakarta Timur, Jakarta 13220, Indonesia

^c Darul Hikmah Islamic College, Bangkalan 69121 Indonesia

Article Info

Article history:

Received 7 June 2024

Revised 13 October 2024

Accepted 18 November 2024

Available online 18 March 2026

Keywords:

elementary school students,
instrument development,
mathematical word problems,
problem-solving skills,
Rasch Analysis

Abstract

This study focused on developing and validating an instrument to assess elementary school students' abilities in solving mathematical word problems, specifically targeting their skills in understanding, planning, applying strategies, and evaluation. Using the ADDIE instructional design model, a 30-item multiple-choice test consisting of word problem-solving was administered to 500 fourth-grade students (252 males, 248 females, aged 9–11) from three schools in Depok, Indonesia. Each of the 30 items was presented in a story-telling question format, requiring students to read and comprehend a real-world mathematical scenario before solving it. The questions were designed to test different cognitive aspects, including the student's ability to understand the problem, plan an appropriate solution, apply mathematical strategies, and evaluate the correctness of their solution. The validity and reliability of the instrument were analyzed using Rasch analysis. Two trials were conducted. In Trial 1 ($n = 200$), the instrument demonstrated high item reliability (0.96) but low person reliability (0.58), with seven items identified as misfitting. Trial 2 ($n = 300$) showed an improvement in person reliability (0.67) and sustained high item reliability (0.98), though overfitting was observed. The results contribute to the creation of a valid and reliable tool for assessing problem-solving abilities in mathematical word problems. The findings emphasize the importance of targeted interventions to enhance students' problem-solving skills, offering insights for improving teaching and assessment practices. This research adds to the development of assessment instruments in mathematical problem-solving and deepens the understanding of elementary students' problem-solving abilities.

© 2026 Kasetsart University.

* Corresponding author.

E-mail address: sepiyanti@gmail.com (Yanti, S.).

† Co-first authors.

Introduction

Mathematics is crucial in modern society, impacting science, technology, economics, and finance (Maass et al., 2019). However, Indonesian students struggle with mathematics, performing poorly in PISA and TIMSS assessments (Ehmke et al., 2020). Specifically, fourth-grade students demonstrate below-average mathematical literacy, struggling with problem-solving and reasoning. Existing assessments lack validity and reliability, hindering accurate evaluations (Putri & Wutsqa, 2019). Limited research focuses on developing context-specific instruments for Indonesian elementary education (Kim et al., 2019). This disparity highlights the need for effective, tailored assessments to identify areas requiring improvement and enhance mathematical word problem-solving skills.

Previous research on mathematical assessment instruments for problem-solving skills has limitations. Studies often lack comprehensive validation methods, such as Rasch analysis, leading to inaccurate measurements (Hellstrand et al., 2020). Instruments typically focus on one aspect of problem-solving, neglecting evaluation and reflection (Alias et al., 2022). Limited, non-representative samples reduce generalizability. Assessments often lack real-world contexts and fail to accommodate individual differences in literacy levels or learning difficulties (O'Neill & Padden, 2021), making them less adaptable to diverse student needs.

Addressing these gaps is essential to improve the accuracy, relevance, and effectiveness of assessments in measuring students' problem-solving abilities. A well-validated instrument ensures that the data collected accurately reflects students' true abilities, leading to better insights for educators and researchers. By incorporating all aspects of problem-solving such as understanding, planning, applying strategies, and evaluating as an instrument can provide a more comprehensive evaluation of students' cognitive skills (Aguhayon et al., 2023). Moreover, using diverse and representative samples enhances the generalizability of the results, making the instrument applicable across different student populations. Real-world contexts in problem-solving tasks make the assessments more meaningful and practical for students, better preparing them for challenges outside the classroom (Nilimaa, 2023). Accounting for individual differences allows the instrument to be more inclusive and adaptive, ensuring that students with various abilities and backgrounds are fairly assessed. Filling these gaps will ultimately

lead to more effective teaching strategies and targeted interventions that can significantly improve student learning outcomes.

This study addresses the gap in mathematical word problem-solving assessments for Indonesian fourth-grade students, aligning with the national curriculum. The instrument evaluates critical skills: understanding, planning, strategy application, and evaluation. Mathematics is vital in daily life, with broad applications. Strong skills are crucial for tackling everyday challenges. Indonesia's mathematics education spans elementary to university levels, ensuring a solid foundation. Superior mathematical abilities correlate with overall education quality (Sumarni et al., 2018). International assessments like PISA and TIMSS evaluate Indonesia's education quality. This study's instrument will enhance assessment validity, supporting effective mathematics education and improved student outcomes.

The mathematical abilities of Indonesian students have consistently ranked among the lowest in the PISA (Program for International Student Assessment) for 15-year-old students, which includes the domains of Mathematics, Science, and Literacy (Setiawan et al., 2022). Since Indonesia's first participation in PISA from 2000 until 2018, the position of Indonesian students has always been at the bottom among participating countries (Rastuti & Prahmana, 2021). Although there was a significant score increase of 11 points in 2015, the ranking remained below the Organisation for Economic Co-operation and Development (OECD) (OECD, 2015) member countries.

The Trends in International Mathematics and Science Study (TIMSS) assesses students' mathematics and science abilities globally. Conducted every four years since 1999, TIMSS provides reliable data for policymakers. Cognitive domains include knowledge, application, and reasoning, with ability levels ranging from advanced to low. Indonesia's TIMSS scores (2003–2015) were consistently below 400, indicating room for improvement. Mathematical problem-solving abilities involve understanding, planning, applying strategies, and evaluating results. Word problems require complex interpretation and resolution skills. Studies show students' problem-solving abilities need enhancement (Santos-Trigo, 2024). Developing mathematical skills through word problems is crucial for daily life applications and problem-solving proficiency.

Literature Review

Research highlights the significance of mathematical word problem-solving skills in elementary education. Studies have explored assessments aligned with Cognitive Load Theory (CLT), reducing extraneous cognitive load and enhancing validity. CLT, developed by John Sweller (1994), emphasizes human working memory limitations and the need for instructional design to facilitate learning. Considering young learners' limited working memory, math word problems must manage cognitive load effectively. Educators should balance intrinsic load (mathematical concept difficulty) and minimize extraneous load (unnecessary complexity) to prevent overwhelm (Chen, 2018; Plass, 2019). Clear, concise wording and step-by-step scaffolding reduce cognitive burden, allocating mental effort toward understanding math concepts (germane load). Applying CLT principles, teachers can design appropriately challenging word problems, facilitating better learning and schema building in elementary mathematics education (Paas, 2020). Effective CLT-based design enhances student outcomes, making math word problems more accessible and enjoyable.

A significant research gap exists in applying Cognitive Load Theory (CLT) to math word problem development for elementary school students. Despite CLT's established relevance in instructional design, its direct implementation in elementary math problem-solving remains underexplored. Previous research provides general guidelines, but lacks specific assessment instruments for elementary math problems (Cook, 2006). Furthermore, limited studies have examined the impact of CLT-informed word problems on learning outcomes in real-world elementary settings. Consequently, several research gaps persist, including the need for practical guidelines on applying CLT principles and a deeper understanding of CLT-informed word problems' effectiveness in enhancing student outcomes.

On the other hand, the ADDIE model has proven effective in developing assessments instrument (Spatioti et al., 2022). Existing research highlights the need for valid assessments, cognitive load management, and problem-solving strategies. Key findings indicate Indonesian students struggle with mathematical literacy, particularly problem-solving and reasoning. Existing assessments lack validity and reliability. This study addresses these gaps by developing a mathematical word problem-solving instrument for Indonesian elementary education, integrating CLT and the ADDIE model.

The instrument aims to assess critical skills, including understanding, planning, strategy application, and evaluation. By enhancing mathematical literacy skills, this study contributes to improving Indonesian students' performance in international assessments like PISA and TIMSS, ultimately benefiting their future academic success (Glaesser, 2019). Effective assessments are crucial for identifying areas requiring improvement.

By applying CLT and the ADDIE model, this study develops an instrument that reduces cognitive load, enhances validity and reliability, and effectively evaluating mathematical word problem-solving abilities. Findings inform the development of an instrument tailored to Indonesian elementary education, addressing gaps in existing assessments. This study addresses a significant research gap by focusing on the development of mathematical word problem-solving instruments for lower-level elementary school students, aligning with the national curriculum and incorporating Cognitive Load Theory (CLT) and the ADDIE model. Unlike previous studies, which primarily concentrated on teaching methods, this research prioritizes instrument development (Paas & Van Merriënboer, 2020).

Methodology

Research Design

This research aimed to construct and validate an assessment instrument to measure and evaluate problem-solving skills in elementary school students by using the ADDIE method, which was modified from (Spatioti et al., 2022). The ultimate goal was to assess the positive impact of the instrument on enhancing students' learning and understanding. The ADDIE model's five stages - Analysis, Design, Development, Implementation, and Evaluation - ensured the production of an effective instrument, as supported by previous studies (Durlak et al., 2011). This systematic approach allowed for comprehensive understanding, design, development, implementation, and evaluation (Meilani et al., 2022; Waisakanitri et al., 2023), ultimately ensuring a positive impact in the educational context (Richter & Richter, 2024).

Data Collection and Sampling

The population focus of this study is limited to elementary school students, specifically those in Depok City, West Java, Indonesia, with a specification

on fourth-grade students in schools with an A accreditation and implementing the Kurikulum Merdeka. The study uses a sample from five elementary schools in Depok City, with a total of 500 respondents as the actual sample. The first trial phase was conducted with a small-scale sample before data collection, followed by the second trial phase using the main sample. The study targeted Indonesian fourth-grade students from public and private elementary schools in Jakarta, Indonesia. The participants were typically 9-10 years old and had completed the third-grade mathematics curriculum. This study employed stratified random sampling to ensure representation from various school types (public and private) and locations (urban and suburban). Schools were selected based on random sampling, and then students were randomly selected from each school. The sample size was calculated using Cochran's formula:

$$n = (Z^2 \times p \times q) / E^2$$

where: n = sample size

$Z = 1.96$ (95% confidence level)

$p = .5$ (estimated proportion of students with adequate mathematical literacy)

$q = 0.5 (1 - p)$

$E = 0.05$ (margin of error)

Substituting the values: $n = (1.96^2 \times 0.5 \times 0.5) / 0.05^2$
 $n = 384.16 \approx 385$

Therefore, the study aimed to recruit 385 Indonesian fourth-grade students.

The sampling procedure involved several steps. First, permission was obtained from school administrators to conduct the study. Next, schools were randomly selected using stratified sampling to ensure representation from various school types (public and private) and locations (urban and suburban). Within these selected schools, students were then randomly chosen to participate. To ensure ethical standards, informed consent was obtained from parents/guardians of the selected students. Finally, the developed mathematical word problem-solving instrument was administered to the selected students, comprising a total sample size of 385 Indonesian fourth-grade students.

Research Instrument Development

The developed assessment instrument consists of word problems aimed at addressing fundamental mathematical issues for elementary school students. It includes a total of 40 questions, with 20 percent categorized as low-order thinking skills (LOTS), 50 percent as medium-order thinking skills (MOTS), and 30 percent as high-order thinking skills (HOTS). The questions are designed

to align with the competency dimensions of students, ranging from levels C3 to C6.

Data Analysis

The results of the instrument development include items and the evaluation of item validity based on the assessment of four experts. The data were recorded in an Excel file and analyzed using Aiken's V validity method. The validity and reliability of the assessment instrument were established through Rasch modeling using Winsteps software, which enables detailed analysis at the item, person, and instrument levels. The following steps for Rasch modeling are 1) item fit order, 2) item difficulty level, 3) differential item functioning (DIF), 4) person ability measures, and 5) measurement information function.

Result

Constructing and validating an assessment instrument is crucial to measure and evaluate problem-solving skills in elementary school students. The following are the results of the construction and validation of the instrument based on the stages of the ADDIE (Analysis, Design, Development, Implementation, and Evaluation) research and development model.

Analysis

The analysis shows that students' difficulties in learning mathematics are influenced by internal factors, such as motivation and intellectual ability, as well as external factors, including teaching methods and family environment. Students often struggle with word problems due to the need for a strong grasp of mathematical concepts and problem-solving skills. Therefore, they require an assessment tool that measures their understanding of math within real-life contexts (Saha et al., 2024). This research phase identified the needs and characteristics of the instrument and developed strategies to address students' challenges with math word problems, aligned with the Kurikulum Merdeka. The curriculum aims to foster independent learning, motivation, creativity, and academic success.

The development of math word problem instruments for fourth-grade students in Depok should include a variety of real-life situations, range in difficulty from simple to complex, and assess cognitive aspects like understanding, application, and analysis. The problems should also be

contextualized in ways that make mathematical concepts relatable to everyday life (Pongsakdi et al., 2019).

Design

At this stage, the research instrument is developed based on the results of the needs analysis conducted in the field. The design process includes selecting the material, context, concepts, and the level and type of questions. This conceptual design serves as the foundation for further development. The analysis phase involves identifying needs based on the curriculum and relevant competency standards for fourth-grade students, as well as understanding their characteristics and learning environments to ensure alignment with their abilities. A review of existing literature and research is conducted to inform the types of math word problems, appropriate difficulty levels, and cognitive aspects to assess. The learning process aims to provide high-quality, interactive experiences relevant to students' lives, using math word problems as assessment tools to evaluate learning objectives. Regular evaluation and revision of the instrument based on feedback from educators and students allow for continuous improvement. Overall, the development of math word problem instruments within the Kurikulum Merdeka framework supports learner-centered education, enabling students to achieve their full potential in mathematics.

Development

The assessment instrument constructed includes a total of 40 questions, thoughtfully categorized to reflect varying levels of cognitive demand. Specifically, 20 percent of the questions are classified as low-order thinking skills (LOTS), which focus on basic recall and comprehension. Meanwhile, 50 percent are categorized as medium-order thinking skills (MOTS), requiring students to apply their knowledge and engage in more

complex reasoning. The remaining 30 percent of the questions are classified as high-order thinking skills (HOTS), which challenge students to analyze, evaluate, and create solutions to intricate problems. Additionally, the development of these questions takes into account the various dimensions of student competency, spanning levels C3 to C6. This ensures that the assessment not only evaluates basic understanding but also encourages deeper cognitive engagement, fostering a more comprehensive development of mathematical skills in young learners.

The mathematical word problem instrument underwent rigorous validation and reliability testing to ensure its accuracy and consistency. Four experts evaluated the instrument's content validity, yielding a high Aiken's V score of 0.81, indicating strong alignment between the measured mathematical content and the developed instrument. Additionally, test-retest procedures demonstrated exceptional reliability with a coefficient of 0.918, confirming the instrument's ability to produce consistent results. Statistical analysis using Pearson Product Moment Correlation at a 0.05 significance level further validated the instrument's reliability. Following expert evaluation, 10 items were eliminated due to low content validity, while 30 items demonstrated moderate to high validity with Aiken's V scores exceeding 0.75. Cronbach's Alpha analysis revealed a remarkable internal consistency score of 0.918, underscoring the instrument's exceptional item reliability. These results collectively establish the instrument's high validity and reliability, instilling confidence in its ability to effectively assess students' mathematical word problem-solving skills. The following is the reliability analysis results based on the Rasch Model (Figure 1):

Reliability is a measure of the trustworthiness of an instrument. The estimated item reliability of 0.90 indicates that the instrument has adequate sensitivity in classifying students based on their abilities (Jang et al., 2022). The reliability data generated from the Winsteps analysis can be seen in the Summary Statistics in Figure 1.

SUMMARY OF 29 MEASURED ITEM								
	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD
MEAN	98.6	200.0	.00	.16	1.00	-.1	1.01	.0
S.D.	36.8	.1	.89	.01	.06	1.2	.08	1.2
MAX.	172.0	200.0	1.50	.21	1.09	1.7	1.14	1.9
MIN.	39.0	199.0	-1.96	.15	.85	-3.0	.82	-2.9
REAL RMSE	.16	TRUE SD	.88	SEPARATION	5.41	ITEM	RELIABILITY	.97
MODEL RMSE	.16	TRUE SD	.88	SEPARATION	5.47	ITEM	RELIABILITY	.97
S.E. OF ITEM MEAN = .17								

Figure 1 Summary Statistic Item

Based on this analysis, the instrument reliability (R) reaches 0.96, indicating a high and dependable level of reliability. Additionally, the item separation index of 5.41 indicates good variation in item difficulty levels. The higher the item separation index, the better the variation in the difficulty levels of the items.

Implementation

The assessment instrument was implemented in five elementary schools in Depok City, with a total of 500 respondents as the actual sample. The first trial phase was conducted with a small-scale sample before data collection, followed by the second trial phase using the main sample. Here are the following implementation results for two phases of implementation.

Trial Phase I

After the mathematical word problem-solving ability instrument has been validated by experts, the next step is to distribute it to respondents, referred to as Trial Phase I or Sample Testing, involving 200 fourth-grade elementary school students in Depok. The data generated from the first trial phase are then analyzed to verify the individual and item validity and reliability using the Rasch Model with Winsteps software. Figure 2 presents the summary results of the Rasch analysis for Trial Phase I, including person-item measures and reliability indices.

Based on Figure 2, the mean measure person value of -0.14 is lower than the mean measure item value, indicating that students' abilities do not match the difficulty levels of the items. Good items are those that measure the same level of ability between items and students. A mean measure < 0.00 indicates a tendency for students to answer fewer items correctly than expected. Rating scale instrument quality criteria with the Rasch model suggests that good person and item measurement reliability values are > 0.67. Based on the analysis above,

the consistency (reliability) of student responses is weak, with a reliability value of 0.58, while the quality of items is exceptional with a reliability value of 0.96.

Item Fit Order

Item Fit refers to the extent to which an item or question is appropriate within the context of measurement or evaluation, indicating how well an item aligns with the intended measurement or evaluation objectives. In item analysis using the Rasch Model, the level of item fit is measured by various statistics, such as Outfit Mean Square (MNSQ) or Infit Mean Square (MNSQ). These values reflect the extent to which participants' or individuals' actual responses to the item align with the predictions of the model or measurement theory used. When the MNSQ value approaches 1, it indicates that the item fits the model well. However, if the MNSQ value is too high or too low, it indicates a mismatch between actual responses and model predictions, which could be caused by factors such as the presence of ambiguous or irrelevant items. Therefore, optimal item fit is crucial to ensure the accuracy and reliability of the measurement or evaluation instrument used in the measurement or evaluation process. Item fit analysis is a key step in the development and validation of evaluation or measurement instruments.

Based on the sum of the Mean and S.D. of Infit Mean Square (MNSQ) = $1.00 + 0.06 = 1.06$. Logit values per item greater than 1.06 categorize the respective items as unfit with the Rasch model. Based on the data above, the unfit items are BS16, BS23, BS14, and BS15 because their Infit MNSQ values are greater than 1.06, namely: 1.07; 1.07; 1.07; 1.09.

Items are considered fit when they consistently behave as expected by the model. According to Boone et al. (2013), three criteria are used: outfit mean square (MNSQ) values, outfit Z-standard (ZSTD) values, and point measure correlation (Pt Mean Corr) values.

PERSON	200 INPUT	200 MEASURED	INFIT		OUTFIT			
	SCORE	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	14.7	30.0	-.14	.43	1.00	.0	1.01	.0
S.D.	3.9	.0	.67	.03	.18	1.2	.27	1.2
REAL RMSE	.43	TRUE SD	.51	SEPARATION	1.18	PERSON	RELIABILITY	.58
ITEM	30 INPUT	30 MEASURED	INFIT		OUTFIT			
	SCORE	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	94.3	200.0	.00	.16	1.00	-.1	1.01	.0
S.D.	33.9	.0	.83	.01	.07	1.5	.10	1.5
REAL RMSE	.16	TRUE SD	.81	SEPARATION	5.03	ITEM	RELIABILITY	.96

Figure 2 Results of Rasch Analysis Trial 1

From these criteria, it is determined that items are considered fit if the MNSQ value falls between $0.5 < \text{MNSQ} < 1.5$, the ZSTD value falls between $-2 < \text{ZSTD} < 2$, and the Point Measure Correlation (Pt Mean Corr) value is non-negative or falls between $0.4 < \text{Pt Measure Corr} < 0.85$.

Item Fit Measure

The difficulty level of items is expressed in statistics such as outfit mean square (MNSQ) or chi-square (χ^2), indicating how much actual student responses to the item differ from what is expected based on the Rasch model. The ideal value for item difficulty level is around 1. Values greater than 1 indicate that the item is too difficult or too easy for the population of students being tested, while values less than 1 indicate that the item is too difficult for that population of students. Therefore, the Item Fit Measure provides an indication of how well the item fits the Rasch model and how accurately it can measure students' abilities.

The difficulty levels of items, can be observed through the Measure values in the corresponding column. A larger Measure value indicates a greater level of difficulty for that item. This finding also indicates that the overall scores move in line with the measure values, reflecting the number of correct responses. For example, for item No. 17, it can be seen that its measure value is 1.50, and only 39 students answered it correctly. In the measured item, useful data can also be provided to classify the difficulty levels of each question (item measure), which can be observed through the standard deviation values (0.89). When this value is combined with the average measure value, there are 4 groups of questions based on their difficulty levels. In this case, the difficulty levels of questions are grouped into 4 types: very difficult, difficult, very easy, and easy. For the difficult question group, there are questions (BS23, BS21, BS29, BS27, BS24, BS28, BS22, BS14, and BS12) with determinations (0.0 logits + 1 SD). Meanwhile, the group of questions with very difficult difficulty levels is found in questions (BS17, BS9, BS15, BS11, and BS10) with determinations (greater than + 1SD). For the grouping of questions with easy difficulty levels, there are questions (BS25, BS26, BS2, BS18, BS16, BS7, BS30, BS19, BS20, and BS8) with determinations (0.0 logits - 1 SD). Conversely, for the grouping of very easy difficulty level questions, they are found in questions (BS13, BS6, BS3, BS5, and BS4) with determinations (less than -1SD).

For the group of most difficult questions, they are located at the top position with a determination greater than +1 logit, comprising 5 test items. Meanwhile, for the group of difficult questions, they are positioned in the middle with determinations (0.0 logit + 1 SD), and for the group of easy questions, they are also in the middle with determinations (0.0 logit - 1 SD). Lastly, for the group of very easy questions, they are located at the bottom position with determinations less than -1 logit.

Differential Item Functioning

The results of the analysis of Differential Item Functioning (DIF) using the Rasch model indicate differences in the functioning of test items among the groups being tested. This can indicate item bias, where there are differences in the likelihood of answering correctly between groups, regardless of ability. These results are important for improving the fairness and accuracy of the test as a whole by addressing or revising items that show item bias. Statistical methods such as Mantel-Haenszel, logistic regression, or item response theory (IRT) are used in DIF analysis. Below are the results of the DIF analysis for each test item obtained from the analysis using the Rasch model.

In this analysis of Differential Item Functioning (DIF), the class specification of DIF is $\text{DIF}=\$S1W1$, indicating that all probability values are greater than 5 percent. From Figure 3, the measurement of Differential Item Functioning (DIF) for a series of test items (BS1 to BS30). In this context, DIF refers to the difference in the probability of answering items correctly between two groups, which in this case may be between "PERSON" and "CLASS". The DIF values listed in the "R" column indicate the extent of the difference between the "PERSON" and "CLASS" groups in the likelihood of answering items correctly. Positive values on BS9, BS10, BS11, BS15, BS17, BS18, BS21, BS22, BS23, BS24, BS27, BS28, BS29 indicate that the "PERSON" group is more likely to answer correctly compared to the "CLASS" group, while negative values on BS1, BS3, BS4, BS5, BS6, BS7, BS8, BS13, BS16, BS19, BS20, BS25, and BS30 indicate the opposite.

Input Uji Coba 2.xlsx

PERSON	300	INPUT	300	MEASURED		INFIT		OUTFIT	
	SCORE	COUNT		MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	18.2	30.0		.78	.48	1.01	.0	.96	.1
S.D.	3.9	.0		.83	.07	.22	1.0	.75	.6
REAL RMSE	.48	TRUE SD		.68	SEPARATION	1.41	PERSON RELIABILITY		.67
ITEM	30	INPUT	30	MEASURED		INFIT		OUTFIT	
MEAN	177.8	300.0		.00	.19	1.00	.0	.96	-.1
S.D.	67.4	.2		1.72	.17	.11	2.2	.22	2.3
REAL RMSE	.25	TRUE SD		1.70	SEPARATION	6.74	ITEM RELIABILITY		.98

Figure 3 Results of Rasch Analysis Trial II

Person Ability Measure

The level of student ability is measured relative to the difficulty of items. The higher the level of student ability, the more likely they are to be able to answer more difficult items in the evaluation instrument. Conversely, students with lower levels of ability may only be able to answer easier items. Individual Ability Level (Person Measure) is typically expressed on a logit scale, where the midpoint of the scale (usually 0) represents the average level of ability in the tested population sample. Positive values indicate ability levels above average, while negative values indicate ability levels below average. Below are the data of 200 students as respondents in test I.

Person Ability is R181 with a logit value of +1.78. This student can correctly answer 25 out of the 30 items in the instrument. Comparatively, students with codes R051, R054, R072, R119, R123, R141, R154, R156, R158, R167, R169, R173, R174, R185, and R199, who have a logit value of +0.09, can correctly answer 16 out of the 30 items.

Students with the lowest abilities are R020 and R021, with a logit value of -1.78, and they can correctly answer 6 out of the 30 items. Looking at the overall abilities of students, the item answered correctly by the most students (20 students) is R006, R033, R109, R118, R121, R122, R131, R144, R149, R151, R160, R163, R164, R171, R176, R183, R184, R187, and R197, with a logit value of +0.41. Students can correctly answer 18 out of the 30 items. Additionally, the average logit person value is known to be -0.14 logit.

For the group of students with high abilities, they are positioned at the top with a determination of (0.0 logit + 1SD), where there are 9 items that can be answered by the most out of 200 student respondents. Meanwhile, for the group of students with low abilities, they are positioned in the middle with a determination of (0.0 logit - 1SD),

where there are 10 items that can be correctly answered by students. Additionally, for the group of students with very low abilities, they are positioned at the bottom with a determination of less than -1 logit, where there are 5 items that can be answered by students. Easy questions are positioned in the middle with a determination of (0.0 logit - 1SD), and very easy questions are positioned at the bottom with a determination of less than -1 logit. From this explanation, it is known that the average level of student ability respondents is above the standard item difficulty level.

Trial Phase II

The summary of Stage II of the instrument test for Mathematics story problems for fourth-grade elementary students involves 300 respondents from 3 schools. The instrument has been revised after Stage I and applied to the selected sample, with data collected according to established procedures. Statistical analysis is conducted to evaluate the reliability, validity, and effectiveness of the instrument.

Figure 3 presents the results of the Rasch analysis for Trial Phase II, including person-item measures and reliability indices. The results of the data analysis are evaluated to determine whether the instrument meets the established objectives and whether any improvements are needed. Key findings, recommendations for improvement, and conclusions are comprehensively presented in the report. This stage is crucial in instrument development as it provides a clearer insight into the quality and performance before widespread use.

Based on Figure 3, the Mean Measure Person value of 0.78 is lower than the mean measure item value, indicating that the student's abilities do not match the difficulty level of the items. Good items are those that measure the same level of ability between the items and the students. A Mean measure < 0.00 indicates a tendency for students

to answer fewer items correctly than expected. The rating scale instrument quality criteria with the Rasch model states that person and item measurement reliability should be > 0.67 . Based on the analysis results, the consistency (reliability) of student responses is moderate with a reliability value of 0.67, while the quality of the items is excellent with a reliability value of 0.98.

Item Fit Order

The item fit, in the context of measurement or evaluation, measures the extent to which an item or question aligns with the established objectives. In Rasch Model item analysis, we use statistics such as Outfit Mean Square (MNSQ) or Infit Mean Square (MNSQ) to measure the level of item fit. These values help us understand how well participants' responses align with the model or theory predictions. The sum of Mean and S.D. Infit Mean Square (MNSQ) is calculated as $1.00 + 0.11 = 1.11$. If the logit value per item exceeds 1.11, the corresponding items are categorized as not fitting the Rasch model. Based on the data above, the items that do not fit are BS21, BS23, BS24, BS28, BS18, BS17, and BS12 because their Infit MNSQ values are greater than 1.11, specifically 1.15, 1.18, 1.14, 1.17, 1.30.

The fit of items is measured based on three criteria established by Boone et al. (2013): the value of outfit mean square (MNSQ), the value of outfit Z-standard (ZSTD), and the value of point measure correlation (Pt Mean Corr). An item is considered fitting if it meets the established criteria, namely, MNSQ value within the range of $0.5 < \text{MNSQ} < 1.5$, ZSTD value within the range of $-2 < \text{ZSTD} < 2$, and Point Measure Correlation (Pt Mean Corr) value is non-negative or within the range of $0.4 < \text{Pt Measure Corr} < 0.85$.

Item Fit Measure

The Item Fit Measure analysis utilizing the Rasch Model in Trial II assessed the conformity of items to Rasch assumptions, ensuring that the probability of correct answers depends solely on student ability. The analysis evaluated item suitability through statistics such as INFIT and OUTFIT Mean Square (MNSQ) and Point Measure Correlation (Pt-Measure). Results revealed varying levels of item suitability, with most items exhibiting acceptable MNSQ and ZSTD values, indicating a good fit to the Rasch model. However, some items displayed discrepancies, warranting further examination.

The average MNSQ and ZSTD values approached the ideal value of 1, indicating adequate overall suitability. Nonetheless, the high standard deviation highlighted significant variation in item suitability. Item measures facilitated the categorization of difficulty levels into four distinct groups: very difficult (10 items, $> +1\text{SD}$), difficult (11 items, $0.0 \text{ logits} +1\text{SD}$), easy (1 item, $0.0 \text{ logits} -1\text{SD}$), and very easy (7 items, $< -1\text{SD}$).

These groupings enable targeted instruction and assessment refinement. Specifically, educators can focus on the most challenging items (BS14, BS10, BS15, BS24, BS22, BS18, BS15, BS23, BS12, and BS17) and adjust instruction to meet students' needs. Conversely, easier items (BS7, BS8, BS6, BS2, BS5, BS4, and BS3) may require minimal instructional emphasis.

Differential Item Functioning

The results of the analysis of Differential Item Functioning (DIF) using the Rasch model revealed variations in how different groups responded to each item. This indicates the potential presence of bias in some items, where differences in the ability to answer items are found among groups, not solely due to differences in their abilities. These findings are crucial in improving the fairness and accuracy of the test overall. Therefore, there is a need to revise or correct items that exhibit such bias. Various statistical methods such as Mantel-Haenszel, logistic regression, or item response theory (IRT) can be employed to analyze DIF. Below are the DIF analysis results for each item, generated from the analysis using the Rasch model.

In this analysis of Differential Item Functioning (DIF), the class specification of DIF is $\text{DIF} = \$\$1\text{W}1$, indicating that all probability values are greater than 5 percent. The curve graph above presents data covering the measurement of DIF (Differential Item Functioning) for a series of test items (BS1 to BS30). In this context, DIF refers to the difference in the probability of answering an item correctly between two groups, potentially between "PERSON" and "CLASS".

The DIF values listed in the "R" column indicate how much difference exists between the "PERSON" and "CLASS" groups in the likelihood of answering an item correctly. Positive values for BS9, BS10, BS11, BS12, BS13, BS14, BS15, BS16, BS17, BS18, BS19, BS20, BS21, BS22, BS23, BS24, BS25, BS26, BS27, BS28, and BS29 indicate that the "PERSON" group is more likely to answer correctly compared to the "CLASS" group, while negative values for BS1, BS2, BS3, BS4, BS5, BS6, BS7, BS8, and BS30 indicate the opposite.

Person Ability Measure

The ability level of students is measured relative to item difficulty, with higher ability levels indicating a greater likelihood of answering challenging items correctly. On a logit scale, where 0 represents average ability, the study of 300 students revealed varying levels of proficiency. The highest ability level was R213 (+3.33 logits), correctly answering 28/30 items, while the lowest was R246 (-1.55 logits), answering only 8 correctly. The average logit person value was 0.78. Ability groups emerged: high ability (+1 logit and above) with 11 items answered correctly, moderate ability (0.0 logit + 1SD) with 10 items, low ability (0.0 logit - 1SD) with 1 item, and very low ability (-1 logit and below) with 8 items. These results highlight the range of abilities among the student population, underscoring the need for tailored instruction and assessment.

Evaluation

Evaluation is conducted at each stage of the research process to ensure that the research objectives are met effectively. This systematic approach allows for continuous assessment and refinement of the methods and strategies employed throughout the study. By incorporating evaluation at each phase, researchers can identify potential challenges and areas for improvement early on, thereby enhancing the overall quality and reliability of the findings. Moreover, this iterative process facilitates adjustments based on feedback and data collected, ensuring that the research remains aligned with its goals. Ultimately, such thorough evaluations contribute to the robustness of the research outcomes, providing valuable insights that can inform future studies and practical applications.

Discussion

Quality of Assessment Instrument Developed

The findings from the analysis and comparison of the first trial with the instrument developed using the ADDIE method highlight significant advancements in the quality of the assessment tool designed to measure mathematical problem-solving abilities. The systematic approach of the ADDIE framework, which includes stages of Analysis, Design, Development, Implementation, and Evaluation, has proven effective in refining the assessment instrument to better meet educational goals (Mayfield, 2011).

The application of the Rasch model further enhanced the reliability of both items and respondents, ensuring that the instrument is capable of effectively differentiating between students with varying levels of ability.

Despite these improvements, the results indicate that personal reliability remains an area requiring further attention. The initial Trial I demonstrated a mean measure person of -0.14 logits and a personal reliability of only 0.58, suggesting inconsistencies in measuring individual student abilities. Research by Boone et al. (2013) emphasizes the importance of ensuring that assessments accurately reflect individual competencies to avoid misrepresentations of student understanding. The subsequent Trial II showed a slight increase in personal reliability to 0.67, which is an improvement but still below the desired level of reliability. This indicates that while progress has been made, ongoing refinements are necessary to enhance the assessment's ability to consistently measure student capabilities.

Item Fit Analysis revealed that several items did not conform to expected fit criteria, with a few items having logits less than the ideal threshold. This finding is consistent with the literature, which states that poorly fitting items can distort the overall assessment results. Additionally, there is the presence of an imbalance in item difficulty levels, some being too easy or too difficult, which suggests a need for recalibrating the items to ensure they align better with the target student population's abilities. While the analysis showed no significant difference in item response based on student characteristics, as indicated by DIF values under 5 percent, it remains critical to monitor item performance across diverse student groups to maintain fairness and validity. Although the development of the assessment instrument using the ADDIE method and Rasch modeling has yielded notable improvements in measuring students' mathematical problem-solving skills, there are still opportunities for refinement. Further analysis of item performance and adjustments to enhance personal reliability will be essential for achieving a more consistent and accurate measurement of student abilities. Future research should focus on iterative testing and evaluation of the instrument to ensure that it remains responsive to the educational needs of students and reflects their true mathematical competencies.

The Instrument Ability to Measure and Evaluate Problem-Solving Skills in Elementary School Students

The ability to solve mathematical word problems is a critical skill for elementary school students, as it not only

reflects their understanding of mathematical concepts but also their ability to apply these concepts to real-world situations. To effectively assess this skill, the development of quality assessment instruments is paramount. Such instruments must be designed with careful consideration of both their psychometric properties and their alignment with educational standards and learning outcomes. The importance of high-quality assessment instruments is supported by various educational theories. For instance, Constructivist Learning Theory posits that students build knowledge through experiences, and assessment should measure their understanding in meaningful contexts. In this light, a well-structured assessment tool that includes mathematical word problems enables educators to gauge students' comprehension and their ability to transfer knowledge to practical scenarios.

Moreover, the Rasch model, a probabilistic measurement theory, offers a robust framework for evaluating the effectiveness of assessment instruments. It provides insights into the reliability and validity of test items by focusing on individual responses and their relation to the underlying construct being measured. The use of the Rasch model in developing instruments to measure problem-solving skills ensures that items are not only reliable but also accurately differentiate between students' abilities. Previous research has highlighted the importance of using such models to enhance assessment quality. In the context of measuring problem-solving skills, prior studies have also identified specific attributes that contribute to the effectiveness of assessment instruments. For instance, research by Jensen et al. (2020) highlighted that assessment items should vary in difficulty and require a range of cognitive skills, from basic understanding to higher-order thinking. This finding aligns with the principles of item development in the ADDIE framework, which advocates for a systematic approach to creating assessment tools that accurately measure a spectrum of abilities.

Conclusion

The analysis and comparison between the first trial and the instrument development using the ADDIE method showed significant improvements in mathematical problem-solving abilities assessment. The developed instrument demonstrated increased reliability of items and persons, enhanced ability to differentiate students with varying abilities, stable Cronbach's alpha values indicating good internal consistency, and improved

separation values for accurate measurement of student abilities. Initial findings revealed students tended to answer incorrectly, with item reliability at 0.96 and personal reliability at 0.58, requiring improvement. Item Fit Analysis identified misfitting items and imbalanced difficulty levels. Following refinements, Trial II showed increased mean measure person (0.78 logits), item reliability (0.98), and personal reliability (0.67). However, Infit and Outfit results indicated overfitting and misfitting items remained. Despite improvements, further refinement is necessary to address misfitting items, enhance personal reliability, and balance difficulty levels for consistent measurement of student abilities.

For future research, it is recommended to investigate the long-term impact of high-quality assessment instruments on students' problem-solving skills and their retention of mathematical concepts. Expanding the research to include diverse student populations will enhance generalizability. Additionally, incorporating technology, such as digital platforms for administering word problems, could increase student engagement and provide real-time performance data. Employing qualitative methods may also yield deeper insights into students' thought processes and strategies for solving mathematical problems, enriching our understanding of their capabilities.

Conflict of Interest

The author declares that there is no conflict of interest.

Reference

- Aguhayon, H., Tingson, R., & Pentang, J. (2023). Addressing Students Learning Gaps in Mathematics through Differentiated Instruction. *International Journal of Educational Management and Development Studies*, 4(1), 69–87. <https://doi.org/10.53378/352967>
- Alias, A., Mohtar, L. E., Ayop, S. K., & Rahim, F. R. (2022). A Systematic Review on Instruments to Assess Critical Thinking & Problem-Solving Skills. *EDUCATUM Journal of Science Mathematics and Technology*, 9(Sp), 38–47. <https://doi.org/10.37134/ejsmt.vol9.sp.5.2022>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2013). Rasch Analysis in the Human Sciences. In *Springer eBooks*. <https://doi.org/10.1007/978-94-007-6857-4>
- Chen, O. (2018). Extending Cognitive Load Theory to Incorporate Working Memory Resource Depletion: Evidence from the Spacing Effect. *Educational Psychology Review*, 30(2), 483–501. <https://doi.org/10.1007/s10648-017-9426-2>
- Cook, M. P. (2006). Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. *Science Education*, 90(6), 1073–1091. <https://doi.org/10.1002/sce.20164>

- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The Impact of Enhancing Students' social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development*, 82(1), 405–432. <https://doi.org/10.1111/j.1467-8624.2010.01564.x>
- Ehmke, T., Van Den Ham, A., Sälzer, C., Heine, J., & Prenzel, M. (2020). Measuring mathematics competence in international and national large scale assessments: Linking PISA and the national educational panel study in Germany. *Studies in Educational Evaluation*, 65, 100847. <https://doi.org/10.1016/j.stueduc.2020.100847>
- Glaesser, J. (2019). Competence in educational theory and practice: a critical discussion. *Oxford Review of Education*, 45(1), 70–85. <https://doi.org/10.1080/03054985.2018.1493987>
- 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/10.1016/j.ijer.2020.101580>
- Jang, Y., Choi, S., & Kim, H. (2022). Development and validation of an instrument to measure undergraduate students' attitudes toward the ethics of artificial intelligence (AT-EAI) and analysis of its difference by gender and experience of AI education. *Education and Information Technologies*, 27(8), 11635–11667. <https://doi.org/10.1007/s10639-022-11086-5>
- Jensen, J. L., McDaniel, M. A., Kummer, T. A., Godoy, P. D. D. M., & St Clair, B. (2020). Testing effect on High-Level Cognitive Skills. *CBE—Life Sciences Education*, 19(3), ar39. <https://doi.org/10.1187/cbe.19-10-0193>
- Kim, S., Raza, M., & Seidman, E. (2019). Improving 21st-century teaching skills: The key to effective 21st-century learners. *Research in Comparative and International Education*, 14(1), 99–117. <https://doi.org/10.1177/1745499919829214>
- Maass, K., Geiger, V., Ariza, M. R., & Goos, M. (2019). The Role of Mathematics in interdisciplinary STEM education. *ZDM*, 51(6), 869–884. <https://doi.org/10.1007/s11858-019-01100-5>
- Mayfield, M. (2011). Creating training and development programs: using the ADDIE method. *Development in Learning Organizations an International Journal*, 25(3), 19–22. <https://doi.org/10.1108/14777281111125363>
- Meilani, M., Risalah, D., & Susiaty, N. U. D. (2022). Pengembangan e-Comic berkarakter terhadap pemahaman matematis siswa pada materi segi empat. *Jurnal Pendidikan Jompa Indonesia*, 1(1), 47–57. <https://doi.org/10.55784/jupenji.vol1.iss1.149>
- Nilimaa, J. (2023). New examination approach for Real-World Creativity and Problem-Solving Skills in Mathematics. *Trends in Higher Education*, 2(3), 477–495. <https://doi.org/10.3390/higheredu2030028>
- O'Neill, G., & Padden, L. (2021). Diversifying assessment methods: Barriers, benefits and enablers. *Innovations in Education and Teaching International*, 59(4), 398–409. <https://doi.org/10.1080/14703297.2021.1880462>
- Organisation for Economic Co-operation and Development. (2015). Concepts and definitions for identifying R&D. In *The measurement of scientific, technological and innovation activities* (pp. 43–79). OECD Publishing. <https://doi.org/10.1787/9789264239012-4-en>
- Paas, F. (2020). Cognitive-Load Theory: Methods to Manage Working Memory Load in the Learning of Complex Tasks. *Current Directions in Psychological Science*, 29(4), 394–398. <https://doi.org/10.1177/0963721420922183>
- Paas, F., & Van Merriënboer, J. J. G. (2020). Cognitive-Load Theory: Methods to manage working memory load in the learning of complex tasks. *Current Directions in Psychological Science*, 29(4), 394–398. <https://doi.org/10.1177/0963721420922183>
- Plass, J. L. (2019). Four Ways of Considering Emotion in Cognitive Load Theory. *Educational Psychology Review*, 31(2), 339–359. <https://doi.org/10.1007/s10648-019-09473-5>
- Pongsakdi, N., Kajamies, A., Veermans, K., Lertola, K., Vauras, M., & Lehtinen, E. (2019). What makes mathematical word problem solving challenging? Exploring the roles of word problem characteristics, text comprehension, and arithmetic skills. *ZDM Mathematics Education*, 52(1), 33–44. <https://doi.org/10.1007/s11858-019-01118-9>
- Putri, A. G. E., & Wutsqa, D. U. (2019). Students' mathematical connection ability in solving real-world problems. *Journal of Physics Conference Series*, 1320(1), 012066. <https://doi.org/10.1088/1742-6596/1320/1/012066>
- Rastuti, M., & Prahmana, R. C. I. (2021). The Programme for International Student Assessment Research in Indonesia. *Jurnal Elemen*, 7(2), 232–253. <https://doi.org/10.29408/jel.v7i2.3289>
- Richter, E., & Richter, D. (2024). Measuring the quality of teacher professional development: A large-scale validation study of an 18-item instrument for daily use. *Studies in Educational Evaluation*, 81(2), 101357. <https://doi.org/10.1016/j.stueduc.2024.101357>
- Saha, M., Islam, S., Akhi, A. A., & Saha, G. (2024). Factors affecting success and failure in higher education mathematics: Students' and teachers' perspectives. *Heliyon*, 10(7), e29173. <https://doi.org/10.1016/j.heliyon.2024.e29173>
- Santos-Trigo, M. (2024b). Problem solving in mathematics education: tracing its foundations and current research-practice trends. *ZDM Mathematics Education*, 56, 211–222. <https://doi.org/10.1007/s11858-024-01578-8>
- Setiawan, A. A., Muhtadi, A., & Hukom, J. (2022). Blended learning and student mathematics ability in Indonesia: A Meta-Analysis Study. *International Journal of Instruction*, 15(2), 905–916. <https://doi.org/10.29333/iji.2022.15249a>
- Spatioti, A. G., Kazanidis, I., & Pange, J. (2022). A comparative study of the ADDIE Instructional Design Model in distance education. *Information*, 13(9), 402. <https://doi.org/10.3390/info13090402>
- Sumarni, W., Supardi, K. I., & Widiarti, N. (2018b). Development of assessment instruments to measure critical thinking skills. *IOP Conference Series Materials Science and Engineering*, 349, 012066. <https://doi.org/10.1088/1757-899x/349/1/012066>
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- Waisakanitri, I. D. A. T., Ganing, N. N., & Wulandari, I. G. A. A. (2023). Media komik digital berbasis problem-based learning muatan IPA (Ekosistem) kelas V Sekolah dasar. *Jurnal Ilmiah Pendidikan Profesi Guru*, 6(1), 57–70. <https://doi.org/10.23887/jippg.v6i1.58651>