

Development of mathematics questions in PISA format with number content

Putri Aisyah¹⁾, Sartika Avrilian²⁾, Tara Nurpadillah³⁾, Natanael Putra Rasjid Ginting⁴⁾, Ratu Ilma Indra Putri⁵⁾, Elsa Susanti^{6)*}, Ruth Helen Simarmata⁷⁾

^{1) 2) 3) 4) 5) 6)* 7)} Program Studi Pendidikan Matematika, Universitas Sriwijaya, Indralaya, Indonesia
**elsasusanti@fkip.unsri.ac.id*

Received: 3 June 2025 | Revised: 29 August 2025 | Accepted: 10 September 2025 | Published Online: 3 December 2025

Abstract

The low level of mathematical literacy among Indonesian students is reflected in the average results of PISA 2022, indicating the need to develop PISA-style test instruments. Therefore, this study aims to develop PISA-based questions on numbers that are valid and reliable. The method used is a development study consisting of preliminary and formative evaluation stages. The preliminary stage includes analyzing original PISA questions, student needs, curriculum, and determining the material. The formative evaluation stage includes self-evaluation, expert review, one-on-one, and small group discussions. The format of the questions developed is two multiple-choice questions and four contextual descriptions equivalent to PISA. Content validation was tested qualitatively through expert comments, while quantitative analysis was conducted after testing the questions on six junior high school students in Palembang who were selected purposively. Quantitative analysis includes item validity, reliability, difficulty level, and discriminative power. The research results indicate that the questions are valid and reliable based on content appropriateness, ease of understanding, smooth implementation, and time efficiency. These questions can be used as a training tool to improve students' mathematical abilities in accordance with PISA standards. However, this study has limitations due to the limited sample size and only reaching a small group, so it is recommended that further research be conducted with a larger sample size and up to the field test stage.

Keywords: Mathematical Literacy; Number Content; PISA; Questions Development

Published by [Linear: Journal of Mathematics Education](#)

This is an open access article under the [CC BY SA](#) license



INTRODUCTION

The Program for International Student Assessment (PISA) is an international assessment conducted every three years by the Organization for Economic Cooperation and Development (OECD). In PISA, the assessment focuses on the reading literacy, science, and mathematics skills of 15-year-old students using real-life contexts and everyday situations (OECD, 2022). PISA serves as an important indicator for evaluating the quality of a country's education system, as its assessment emphasizes students' ability to apply knowledge and skills in real-life situations. Indonesia is one of the countries participating in this international assessment.

However, Indonesia's results are still relatively low, especially in mathematics literacy. In fact, the quality of education in Indonesia is still very low and needs improvement, especially in mathematics (Pereira et al., 2022). This is reflected in the latest PISA results, which show that Indonesia has not yet achieved optimal results.

The PISA results announced on December 5, 2023, show that Indonesia ranks 68th out of 81 countries worldwide, with a score of 379 in mathematics, 398 in science, and 371 in reading (OECD, 2023). Compared to the OECD average, the scores obtained are still relatively low, especially in mathematical literacy. PISA results show that the quality of Indonesian students' mathematical literacy is still low (Elyasarikh & Masriyah, 2024).

This cannot be blamed; it is a collective correction that the mathematics questions in the PISA study test argumentation, logic, and problem-solving skills rather than memorization and computation skills. Meanwhile, several studies conducted in various schools in Indonesia show that students are still not accustomed to questions that require logical reasoning to solve real-world problems. According to Mahmudah (2018), Indonesian students still have low mastery of the material and difficulty in answering questions that require reasoning. Theoretical and procedural answers are still preferred and accepted by students. As a result, learning must become accustomed to solving problems that require logical reasoning. This should be a top priority in Indonesia's future education programs (Afriyanti et al., 2018).

To teach and improve mathematical skills in everyday life, PISA-style questions can be used as an alternative learning assessment tool. PISA-style evaluation tools can also help students improve their higher-order thinking skills. The characteristics of PISA questions can be used to determine the extent of students' abilities and whether they fall into the Higher-Order Thinking or Lower-Order Thinking categories. To measure the extent to which learning process competency standards are met, evaluation tools are needed (Herman et al., 2022). Previously, PISA questions researched by other researchers used numerical content. Numerical content is one of the foundational elements of mathematical literacy and consistently appears in PISA questions every year. This content plays a crucial role because it underlies many aspects of daily life, such as transactions, price calculations, and currency conversions, while also being an area where Indonesian students tend to struggle in answering PISA questions (Putra et al., 2016). Therefore, the development of PISA questions based on numerical content was chosen in this study because it is considered relevant, fundamental, and necessary for building students' numeracy skills.

The need for question development is also in line with the direction of the Merdeka Curriculum implemented in Indonesia. This curriculum emphasizes the importance of

strengthening critical thinking, problem solving, and the application of science in real life. The development of PISA questions based on number content can be one of the tools to support the achievement of curriculum objectives, especially in improving applied and contextual mathematical literacy skills.

Written and non-written assessments can be used to measure individual mathematical reasoning abilities. However, there are not many questions designed to measure students' mathematical thinking abilities in Indonesia. Since national exam questions are multiple-choice, students can answer by gue (Ma'rifah & Kristiana, 2015). Based on this, an evaluation tool to measure reasoning abilities must be created.

Although several studies have been conducted on the development of PISA questions, most have not systematically presented a complete formative evaluation process or focused specifically on number content with a local context that is familiar to students. There are still shortcomings in terms of comprehensive content validation and reliable testing based on field data. This study aims to bridge this gap. The PISA mathematics questions on number content are referenced in the evaluation instruments created for this study, which aim to be potentially effective, valid, and reliable.

METHODS

We use the formative evaluation, which was popularized by a philosopher of science named Michael Scriven in his book entitled "The Methodology Of Evaluation" in 1967. The initial steps and formative evaluation stage, which include self-evaluation, prototyping (expert review, one-to-one, and small group), are part of this research, which is a development study (Amalia et al., 2020).

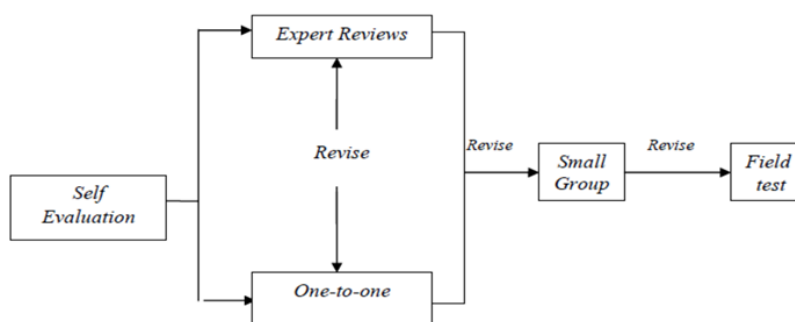


Figure 1. Formative Evaluation Design Flow

The PISA mathematics questions were first reviewed by researchers, who then translated them into Indonesian. After that, a formative evaluation was conducted, with self-evaluation as the first step. The researchers themselves assessed the evaluation instruments that had been developed. Prototype I was the result of this self-evaluation. This research was conducted to

produce useful and valid questions. In parallel, expert review and one-to-one interviews were conducted to validate Prototype I. Experts evaluated the items in Prototype I during the expert review stage. The next validation stage was the validity of the items in Prototype I through one-to-one interviews with three junior high school students with heterogeneous abilities. The result of the revision of Prototype I was Prototype II. Next, Prototype II was tested on students in a small group setting. The questions were administered to six students with high, moderate, and low abilities at one junior high school in Palembang during the 2024/2025 academic year to assess the practicality of the developed questions. The researcher only reached the small group stage and did not proceed to the field test stage.

In this study, walkthroughs, observations, and tests were the methods used for data collection and analysis for context, content and level. Based on the recommendations and comments from the expert review, walkthroughs were used to assess the validity of the evaluation instruments based on language, construct, and substance (N. W. Saputri et al., 2020). The expert review in this study was conducted by mathematics education lecturers from Sriwijaya University. During the initial data analysis, observations were conducted to determine the needs and characteristics of the students. Expert review provides an assessment of the six questions that have been developed qualitatively by providing comments. The assessment results are then revised and a small group discussion is held to obtain the validity of the questions. When the students were given the test questions, observations were also conducted. After the students completed the questions, tests or trials were conducted to determine the practicality of the questions. The question trials were conducted on six students from one junior high school in Palembang with heterogeneous abilities (high, medium, low). To determine the validity and reliability of the questions, the researchers analyzed student responses using quantitative methods. Item validity was assessed using Pearson product-moment correlation between each item's score and the total test score, following the guidelines of Gronlund and Linn (1990). Reliability was evaluated using Cronbach's Alpha. Difficulty level was defined as the proportion of students who answered correctly, and discriminating power was determined by comparing the performance between the top 27% of the group and the bottom 27% of the group. To determine the validity level of each item, it is necessary to interpret the correlation coefficients obtained (Putri, 2021). These correlation coefficient values then form the basis for classifying item validity into the categories listed in Table 1. The reliability categories in this study are based on the classification by Novia, Wardani, Canda, Nurdi & Nurmasiyat (2020). This classification is used to interpret the calculated reliability coefficient score in Table 2.

To determine the validity level of each item, it is necessary to interpret the correlation coefficients obtained (Putri, 2021). These correlation coefficient values then form the basis for classifying item validity into the categories listed in Table 1.

Table 1. Classification of Test Validity Coefficients

Scale Criteria	Category
0.80 – 1.00	Very High
0.60 – 0.80	High
0.40 – 0.60	Enough
0.20 – 0.40	Low
0.00 – 0.20	Very Low

Table 2. Classification of Test Reliability

Scale Criteria	Category
0.81 – 1.00	Very High
0.61 – 0.80	High
0.41 – 0.60	Enough
0.21 – 0.40	Low
0.00 – 0.20	Very Low

RESULTS AND DISCUSSION

Preliminary Stage

In the preliminary stage, preparations were made by analyzing the original PISA questions, analyzing student needs, analyzing the curriculum, and determining the material for the questions to be developed. The PISA questions analyzed by the researchers were those from 2003, 2006, 2009, 2012, 2015, 2018, and 2022. Subsequently, the PISA questions that had been previously analyzed and selected by the researchers were developed further, with each question having a different context.

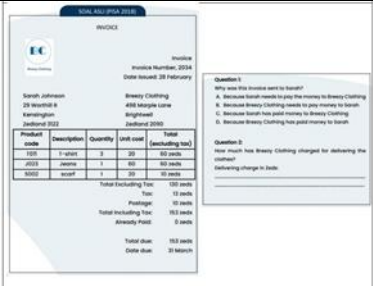
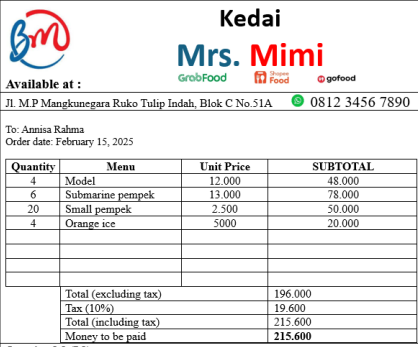
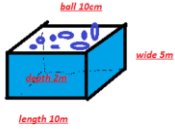
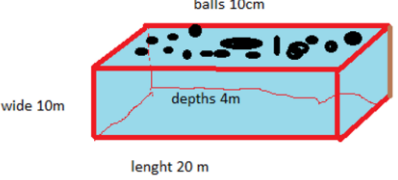
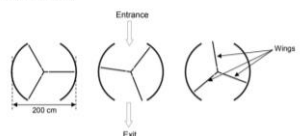
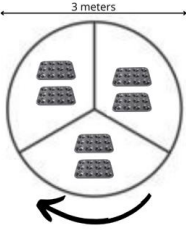
Formative Evaluation

The stages carried out after the preliminary stage are the formative evaluation stage, which consists of self-evaluation, expert review, one-to-one, and small group stages.

Self Evaluation

The first stage of formative evaluation is self-evaluation, where researchers begin to develop PISA-style questions in accordance with the PISA framework. The original PISA questions obtained are adapted to align with the level of each question, ranging from level 1 to level 6, as well as their context and content. The PISA questions developed by the researcher are referred to as prototype 1. The Table 1 compares the original PISA questions with those developed by the researcher.

Table 3. Comparison of Original PISA Questions and Prototype I

Original PISA Questions	Year of PISA Questions	PISA Development Questions (Prototype I)
	2018	<p>1. Annisa ordered several menus at Kedai Bu Mimi, and received the following receipt:</p> 
<p>Question 3: EXCHANGE RATE During these 3 months the exchange rate had changed from 4.2 to 4.0 ZAR per SGD. Was it in Mrs Ling's favour that the exchange rate now was 4.0 ZAR instead of 4.2 ZAR, when she changed her South African rand back to Singapore dollars? Give an explanation to support your answer.</p>	2003	<p>Question 1.1 (L1): Why did Annisa receive this receipt from Kedai Bu Mimi?</p> <p>Question 1.2 (L2): How much money does Annisa have to pay for the food and drinks alone (excluding tax)?</p> <p>Question 1.3 (L3): Kedai Bu Mimi is having a "5% Discount" promo for orders over Rp200,000 (excluding tax). What is the total bill (including tax) that Annisa has to pay if she adds one more order of orange ice?</p> <p>Over the past 6 months, the exchange rate between the Indonesian rupiah (IDR) and the US dollar (USD) has changed from 15,500 IDR per USD to 15,200 IDR per USD. Does the change in the exchange rate to 15,200 IDR per USD benefit or harm Budi if he wants to exchange Indonesian rupiah back into US dollars? Explain your reasoning with calculations that support your answer!</p>
 <p>A rectangular pool is 10 meters long, 5 meters wide, and has an average depth of 2 meters. The pool is filled with small plastic balls that float on the surface of the water. Each ball has a diameter of 10 cm.</p> <p>Question: Estimate how many balls are needed to cover the entire surface of the pool without overlapping.</p>	2000	 <p>There was a father who owned a rectangular-shaped reservoir that was 20 meters long, 10 meters wide, and an average depth of 4 meters. The pond is filled with shade balls that float on the surface of the water. Each ball has a diameter of 10 cm.</p> <p>Question: how many balls are needed to cover the entire pool surface without overlapping</p>
<p>REVOLVING DOOR</p> <p>A revolving door includes three wings which rotate within a circular space. The inside diameter of this space is 2 meters (200 centimeters). The three door wings divide the space into three equal sectors. The plan below shows the door wings in three different positions viewed from the top.</p>  <p>Question 3: REVOLVING DOOR</p> <p>Question intent: Quantity</p> <p>The door makes 4 complete rotations in a minute. There is room for a maximum of two people in each of the three door sectors.</p> <p>What is the maximum number of people that can enter the building through the door in 30 minutes?</p>	2012	<p>A bakery uses an automated machine to produce donuts. This machine has three rotating arms that work in a circular space with a diameter of 3 meters. Each rotating arm divides the circular space into three equal areas. The machine rotates 5 times per minute, and each area can hold a maximum of 2 donut trays. Each tray can produce 12 donuts.</p>  <p>The factory plans to run this machine for 1 hour non-stop to fulfill a large order. Each donut is sold for Rp 5,000, and the production cost per donut is Rp 2,000.</p> <p>How many donuts can the machine produce in 1 hour, and how much profit will the factory earn from selling these donuts?</p>

Expert Review and One-to-One Stages

Next, to improve prototype I, validation was carried out by experts in an expert review with the aim of assessing the validity of the questions developed, namely in terms of content, construct, and language. The characteristics assessed in terms of content were the suitability of

the content quantity in PISA. From the construct perspective, the focus was on the alignment of the developed questions with various contexts and their adaptation to students' ability levels, emphasizing students' ability to analyze problems in line with real-world scenarios that are familiar to them. From the language perspective, the focus is on the alignment of the language used with the Improved Indonesian Spelling Rules, using sentences that are easy to understand, and ensuring there is no ambiguity in the interpretation of sentences.

In line with the expert review validation, one-to-one trials were also conducted with the aim of observing and obtaining feedback from students regarding the clarity of the language and the difficulties students encountered in completing each question. The results of the expert review validation can be seen in Table 4. Meanwhile, the results of the one-to-one validation can be seen in Table 5.

Table 4. Expert Review Validation Results

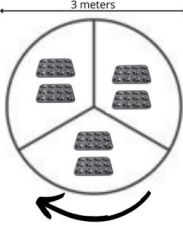
Level	Comments/Suggestions
1 and 2	For multiple choice numbers 1 and 2, the choice (abcd) should be changed to capital letters (ABCD) according to PISA multiple choice questions in general.
2	For option number 2, it should be changed according to the smallest order.
4	For question 4, the value of the rupiah against the US dollar must be in line with the facts.
5	For question 5, the context of the original PISA question and the developed question are different. Validator 2 suggested changing only the size and price.
6	<p>For question 6, the context of “Donut Factory” does not match the facts, so the validator suggests changing it to the actual facts.</p> <div data-bbox="496 1317 1066 1798" style="border: 1px solid black; padding: 5px;"> <p>A bakery uses an automated machine to produce donuts. This machine has three rotating arms that work in a circular space with a diameter of 3 meters. Each rotating arm divides the circular space into three equal areas. The machine rotates 5 times per minute, and each area can hold a maximum of 2 donut trays. Each tray can produce 12 donuts.</p>  <p>The factory plans to run this machine for 1 hour non-stop to fulfill a large order. Each donut is sold for Rp 5,000, and the production cost per donut is Rp 2,000.</p> <p>How many donuts can the machine produce in 1 hour, and how much profit will the factory earn from selling these donuts?</p> </div>

Table 5. One-to-one Validation Results

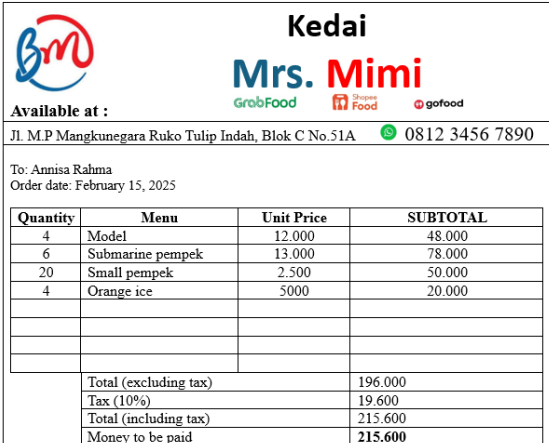

Level	Comments/Suggestions
5	OtO 1: For question 5, the question given is unclear because it does not have a question. Also, the image provided is unclear, so it is difficult to understand the meaning of the question. OtO 2: For question 5, the question provided insufficient information, making it difficult to answer. OtO 3: For question 5, the question is incomplete. It would be better to provide a more complete and clear question to make it easier to understand.
6	OtO 1: For question 6, the illustration provided is unclear, making it difficult to understand. OtO 2: For question 6, it does not mention when the donuts are cooked, and the image is still difficult to imagine. OtO 3: For question 6, there is still insufficient information regarding how many minutes it takes for the donuts to cook.

As a follow-up to the expert review and one-to-one process, six mathematics questions were refined to better align with the characteristics of PISA-based mathematics literacy questions. The refinements covered various aspects, including content, construction, and language.

From the two stages described above, the researchers revised the wording of several questions in accordance with general Indonesian spelling guidelines, changed unclear information in the questions and images, and corrected the context in line with the facts. In the one-to-one stage, it was found that students' ability to read questions and interpret their meaning in mathematical questions was quite good. However, at levels 5 and 6, students had difficulty reading questions because the questions presented were difficult to understand. This resulted in no students being able to identify the questions given and connect their solutions to the material. This serves as feedback for researchers to clarify the text of the questions and images so that students can more easily understand the meaning of the questions. This indicates that students at the one-to-one stage generally understand PISA-style math questions, although there are some questions at levels 5 and 6 that still need to be addressed.

The results of the revisions from the Expert Review and also one to one, also known as Prototype II, can be seen in comparison with Prototype 1 in Table 6.

Table 6. Comparison of Prototype I and Prototype II

Prototype I	Prototype II
<p>1. Annisa ordered several menus at Kedai Bu Mimi, and received the following receipt:</p> <div style="border: 1px solid black; padding: 5px;">  </div> <p>Question 1.1 (L1): Why did Annisa receive this receipt from Kedai Bu Mimi?</p> <p>Question 1.2 (L2): How much money does Annisa have to pay for the food and drinks alone (excluding tax)?</p> <p>Question 1.3 (L3): Kedai Bu Mimi is having a "5% Discount" promo for orders over Rp200,000 (excluding tax). What is the total bill (including tax) that Annisa has to pay if she adds one more order of orange ice?</p>	<p>1. Annisa ordered several menus at Kedai Bu Mimi, and received the following receipt:</p> <div style="border: 1px solid black; padding: 5px;">  </div> <p>Question 1.1 (L1): Why did Annisa receive this receipt from Kedai Bu Mimi? A. Because she needs to pay the bill to Kedai Bu Mimi. B. Because Annisa needs to get a receipt to take home C. Because Bu Mimi needs to pay the bill to Annisa D. Because Mrs. Mimi needs to get a receipt to take home</p> <p>Question 1.2 (L2): How much money should Annisa pay for her food and drinks only (excluding tax)? A. 19.600 B. 32.500 C. 196.000 D. 215.600</p> <p>Question 1.3 (L3): Kedai Bu Mimi is having a "5% Discount" promo for orders over Rp200,000 (excluding tax). What is the total bill (including tax) that Annisa has to pay if she adds one more order of orange ice?</p>

Before the revision, questions 1 and 2 were short answer questions.

Over the past 6 months, the exchange rate between the Indonesian rupiah (IDR) and the US dollar (USD) has changed from 15,500 IDR per USD to 15,200 IDR per USD.

Does the change in the exchange rate to 15,200 IDR per USD benefit or harm Budi if he wants to exchange Indonesian rupiah back into US dollars? Explain your reasoning with calculations that support your answer!

Before the revision, the exchange rate figures did not reflect the actual situation.

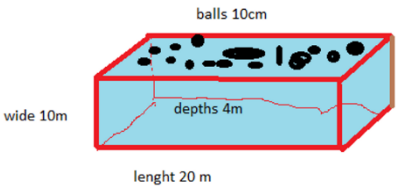
After revision, questions 1 and 2 were changed to multiple choice format.

Over the past 6 months, the exchange rate between the Indonesian rupiah (IDR) and the U dollar (USD) has changed from 15,475 IDR per USD to 16,580 IDR per USD.

Does the change in the exchange rate to 16,580 IDR per USD benefit or harm Budi if he wants to exchange Indonesian rupiah back into US dollars? Explain your reasoning with calculations that support your answer!!

After revision, the exchange rate figures are difficult to change in accordance with the facts.

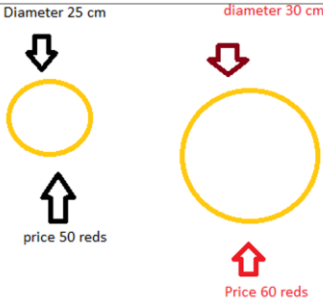
After revision, the questions have been changed to match the original PISA questions.



There was a father who owned a rectangular-shaped reservoir that was 20 meters long, 10 meters wide, and an average depth of 4 meters. The pond is filled with shade balls that float on the surface of the water. Each ball has a diameter of 10 cm.

Question: how many balls are needed to cover the entire pool surface without overlapping

Before revision, the questions on development were not in line with the original PISA questions.



There was a child who wanted to buy a pizza for his parents, the child was offered two round pizzas with the same thickness but different sizes. The small pizza had a diameter 25 cm and was sold for 50 reds, while the large pizza had a diameter of 30 cm and was priced at 60 reds.

Question: Which pizza provides cheaper value for money?

After revision, the questions are now consistent with the original PISA questions.

Prototype I

A bakery uses an automated machine to produce donuts. This machine has three rotating arms that work in a circular space with a diameter of 3 meters. Each rotating arm divides the circular space into three equal areas. The machine rotates 5 times per minute, and each area can hold a maximum of 2 donut trays. Each tray can produce 12 donuts.



The factory plans to run this machine for 1 hour non-stop to fulfill a large order. Each donut is sold for Rp 5,000, and the production cost per donut is Rp 2,000.

How many donuts can the machine produce in 1 hour, and how much profit will the factory earn from selling these donuts?

Prototype II

A donut factory uses an automatic machine to make donuts. This machine has four mold each holding two donuts. After the donuts are molded, they are immediately fried. It takes 90 seconds to fry eight donuts. Each donut is sold for Rp 5,000, and the production cost per donut is Rp 3,500. The factory plans to run the machine continuously until it fulfills an order of 2,080 donuts. How much time is spent producing these donuts, and what is the profit earned by the factory from the sale of these donuts?



<https://www.sccbakery.com/product/brand/all-products/gas-automatic-donut-fryer>

Before the revision, Prototype I still used a manually made automatic donut machine and some of the information in the questions was not contextual.

After revision, Prototype II now uses images and information that are contextually appropriate.

Small Group

After revising the results of the expert review and one-to-one review, also known as prototype II, the next step was to conduct a small group trial of prototype II questions with six heterogeneous students at a public junior high school in Palembang. The purpose of this trial is to assess the practicality of the PISA questions with numerical content that have been developed. In line with the research (Nugraha et al., 2025) that conducted a trial on a small group of six heterogeneous students to determine the practicality of the questions. The items demonstrated appropriate content,

- 1) Were easy for students to understand,
- 2) Could be implemented smoothly,
- 3) And could be completed within the allocated time.

These findings suggest that the developed questions are feasible for use in real classroom settings. The trial was conducted by six students offline for 45 minutes.

Based on the results of the students' answers during the small group session, the validity of each item was tested, where $r_{xy} > r_{tabel}$ with $r_{tabel}(\alpha = 5\%, df = 4) = 0,8114$ therefore, all items developed by the researcher are considered valid with the values in Table 5. Validity testing is a test used to determine whether a measurement tool is valid (reliable) or invalid (H. A. Saputri et al., 2023).

Table 7. Validity Test

Level	Validity	Description
1	0,9713	Very High
2	0,9713	Very High
3	0,8534	Very High
4	0,9486	Very High
5	0,8287	Very High
6	0,9474	Very High

After testing the validity of the questions with results of $r_{xy} > r_{tabel}$, the researchers continued with reliability testing. According to (Magdalena et al., 2023), validity and reliability tests in education are important. For multiple-choice questions, the reliability level of the instrument is 1, while for essay questions, the reliability level of the instrument is 0.8931. The results show a very high reliability value, as seen in Table 7, meaning that this research instrument is consistent and reliable. This aligns with the research by (H. A. Saputri et al., 2023), which states that the level of reliability is empirically indicated by a figure called the reliability coefficient, ranging from 0 to 1, where a higher reliability figure indicates more consistent measurement results (Farida & Musyarofah, 2021).

Table 8. Reliability Test

Type	Level	r_{11}	Category
Multiple Choice	1 & 2	1	Very high reliability
Essay	3, 4, 5, & 6	0,8931	Very high reliability

In addition to validity and reliability tests, measuring the level of difficulty also determines the quality of a good question (Zuhri et al., 2024). From the calculation of difficulty levels, the results varied, with 2 easy questions in the multiple-choice section and 2 moderate and 2 difficult questions in the essay section, as shown in Tables 9 and 10, indicating that the questions developed by the researchers meet balanced criteria. As stated by (Nurhalimah et al., 2022), a good test not only meets validity and reliability criteria but also has a balanced level of difficulty in its questions.

Next is to test the distractors in multiple-choice questions. A distractor is considered good if many students choose it compared to the answer that matches the key (Mustaqim & Sulisti, 2024). From the distractor analysis calculations, based on the Omitted (O) analysis, both items are considered valid, as can be seen in Table 9.

The final stage of testing is to test the Discrimination Power (DP) on the developed questions, which aims to distinguish between high-ability students and low-ability students (Mustaqim & Sulisti, 2024). From the discrimination power calculations, two multiple-choice

questions were found to be of sufficient quality, and four essay questions were found to be of very good quality, as shown in Tables 10 and 11.

Table 9. Multiple Choice Difficulty Level and Excerpt Analysis

Level	Difficulty Level			Question Categories	Excerpt Analysis
	<i>N</i>	<i>B</i>	$\frac{B}{N}$		Description
1	6	5	0,83	Easy	Good
2	6	5	0,83	Easy	Good

Table 10. Level Of Difficulty Test and Power of The Essay

Level	Level of Difficulty		Discrimination Power		
	Difficulty	Description	Calculation Results	Discriminant Power Coefficient	Description
3	0,466	Moderate	$3,66 - 1 = 2,66: 5 = 0,53$	$0,53 \geq 0,40$	Very good
4	0,266	Difficult	$2,66 - 0 = 2,66: 5 = 0,53$	$0,53 \geq 0,40$	Very good
5	0,366	Moderate	$3,66 - 0 = 3,66: 5 = 0,73$	$0,73 \geq 0,40$	Very good
6	0,266	Difficult	$2,66 - 0 = 2,66: 5 = 0,53$	$0,53 \geq 0,40$	Very good

Table 11. Multiple Choice Discrimination Power

Level	Calculation Results	Discriminant Power Coefficient	Description
1	$1 - 0,66 = 0,33$	$0,20 \leq 0,33 < 0,40$	Enough
2	$1 - 0,66 = 0,33$	$0,20 \leq 0,33 < 0,40$	Enough

PISA questions are developed using everyday contexts. This is in line with the statement by Baka, Laksana & Dhiu (2019) that learning should start with something familiar to students. Similarly, Rakhmawati & Alifia (2018) state that for mathematics to be meaningful to humans, it must be relevant to human life, familiar to students, and connected to reality. It is important to instill mathematical content as a human activity. Developing PISA questions in the context of everyday life, such as traditional food, modern food, currency exchange, and food production processes, can attract students' interest and give them the impression that these issues often occur in their lives because they frequently encounter them around them.

Based on the students' answers, it appears that the questions developed can bring out the students' interpretation skills. Students were able to interpret the first question by understanding the purpose of the receipt provided. In answering the second question, students were able to determine the amount of money that had to be paid under certain conditions. Based on the answers to the third question, students were able to apply the question by determining the total

bill with the applicable discount. Then, based on the answers to the fourth question, students were able to analyze changes in currency exchange rates and their impact on the amount of money obtained in other currencies. In answering the fifth question, students were able to evaluate which price was cheaper in order to save money. Finally, based on the answers to the sixth question, students were able to predict the question by combining mathematical concepts with real life to calculate the total donut production and profit obtained. Students' ability to interpret and solve questions developed in accordance with everyday situations using their mathematical concepts demonstrates that they have mathematical literacy skills. The OECD (2019) states that mathematical literacy is the ability to reason mathematically and to formulate, use, and interpret mathematics to solve problems related to everyday life. Students' ability to interpret receipts, calculate payments with conditions, apply discount concepts, analyze currency exchange rates, evaluate price comparisons, and predict business profits demonstrates mastery of these three cognitive processes in real-life contexts.

However, this study has limitations in terms of generalizing the results because it only involved a limited sample from a specific school. Additionally, the assessment of mathematical literacy skills in this study was limited to six contextual questions, so it cannot fully describe all aspects of students' mathematical literacy skills according to the PISA framework.

CONCLUSION

This study successfully developed PISA-based mathematics questions with number content for junior high school students through preliminary and formative evaluation stages, resulting in valid and reliable instruments in terms of content, construct, and language. The questions developed based on PISA questions from 2000 to 2022 are capable of measuring higher-order thinking skills in everyday life contexts and can be utilized by teachers as a reference for more contextual evaluations, as well as to prepare students for international tests such as PISA. However, this study has limitations in that the content is restricted to number-related material and has not yet reached the field test stage. Therefore, further research is needed to develop questions with more diverse content and contexts and conduct field tests to assess the effectiveness of the questions in actual learning, thereby providing a more comprehensive impact on improving students' mathematical literacy.

REFERENCES

Afriyanti, I., Wardono, & Kartono. (2018). Pengembangan Literasi Matematika Mengacu PISA Melalui Pembelajaran Abad Ke-21 Berbasis Teknologi. *PRISMA, Prosiding Seminar Nasional Matematika, 1*, 608–617.

- Amalia, W., Mulyono, & Napitupulu, E. E. (2020). Development of Pisa-Like Mathematical Problems on The Change and Relationship Content to Measure the Mathematical Solution Ability of Middle School Students Development of Pisa-Like Mathematical Problems on The Change and Relationship Content to Measure t. *International Journal of Advanced Science and Technology*, 29(6), 4841–4849.
- Baka, N. A., Laksana, D. N. L., & Dhiu, K. D. (2019). Konten Dan Konteks Budaya Lokal Ngada Sebagai Bahan Ajar Tematik Di Sekolah Dasar. *Journal of Education Technology*, 2(2), 46. <https://doi.org/10.23887/jet.v2i2.16181>
- Elyasarikh, A. A., & Masriyah, M. (2024). Bagaimana Literasi Matematis Siswa pada Penyelesaian Soal PISA-Like Berdasarkan Tingkat Kecerdasan Logis Matematis? *MATHEdunesa*, 13(2), 451–467. <https://doi.org/10.26740/mathedunesa.v13n2.p451-467>
- Farida, & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'Arrib: Journal of Arabic Education*, 1(1), 34–44. <https://doi.org/10.32923/al-muarrib.v1i1.2100>
- Herman, T., Hasanah, A., Nugraha, R. C., Harningsih, E., Ghassani, D. A., & Marasabessy, R. (2022). Pembelajaran Berbasis Masalah-High Order Thinking Skill (HOTS) pada Materi Translasi. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, 6(1), 1131–1150. <https://doi.org/10.31004/cendekia.v6i1.1276>
- Ma'rifah, F., & Kristiana, A. I. (2015). *Pengembangan Paket Tes Kemampuan Penalaran Matematika Model PISA Konten Quantity Pada Siswa Kelas VIII SMP (Development Of Mathematical Reasoning Test Package PISA Model Quantity Contents For Grad 8 th Of Junior High School)*. 1–6.
- Magdalena, I., Fitroh, A., Fadhilah, D., Habsah, D., & Qodrawati, R. (2023). Mengelola Data Uji Validitas dan Reliabilitas dalam Penelitian Pendidikan: Instrumen Tes dan Non Tes Peserta Didik Kelas IV SDNPondok Kacang Barat 03. *Jurnal Pendidikan Sosial Dan Konseling*, 1(2), 49–53. <https://jurnal.ittc.web.id/index.php/jpdsk/article/view/18>
- Mahmudah, W. (2018). Analysis of Student Errors in Solving Hots Type Math Problems Based on Newman's Theory. *Jurnal UJMC*, 4(1), 49–56.
- Mustaqim, M., & Sulisti, H. (2024). Analisis Butir Soal Pas Matematika Peminatan: Daya Pembeda, Tingkat Kesukaran, Dan Kualitas Pengecoh. *Al-'Adad: Jurnal Tadris Matematika*, 3(1), 44–56. <https://doi.org/10.24260/add.v3i1.3011>
- Novia, T., Wardani, A., Canda, C., Nurdi, N., & Nurmasyitah, N. (2020). Analisis Validitas dan Reliabilitas Butir Soal UTS Fisika Kelas X SMA Swasta Muhammadiyah 4 Langsa. *GRAVITASI: Jurnal Pendidikan Fisika Dan Sains*, 3(01), 19–22. <https://doi.org/10.33059/gravitasi.jpfs.v3i01.2256>
- Nugraha, A., Meika, I., & Yunitasari, I. (2025). *Pengembangan Soal High Order Thinking Skill Matematika SMP*. 10(2), 188–194. <https://doi.org/10.30653/003.2024102.366>
- Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). Hubungan Antara Validitas Item Dengan Daya Pembeda Dan Tingkat Kesukaran Soal Pilihan Ganda Pas. *Natural Science Education Research*, 4(3), 249–257. <https://doi.org/10.21107/nser.v4i3.8682>
- OECD. (2019). *PISA 2018 insights and interpretations*. OECD Publishing.
- OECD. (2022). *PISA 2022 Results (Volume I and II) - Country Notes: Indonesia*. FACTSHEETS - INDONESIA. <https://www.oecd.org/en/publications/pisa-2022-results->

volume-i-and-ii-country-notes_ed6fbcc5-en/indonesia_c2e1ae0e-en.html

- OECD. (2023). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing. <https://doi.org/https://doi.org/10.1787/53f23881-en>
- Pereira, J., Aulingga, A., Ning, Y., & Vilela, A. (2022). Kesalahan Siswa Smp Dalam Menyelesaikan Soal Pisa Konten Space and Shape Berdasarkan Teori Newman. *JPMI (Jurnal Pembelajaran Matematika Inovatif)*, 5(2), 317. <https://doi.org/10.22460/jpmi.v5i2.9910>
- Putra, Y. Y., Zulkardi, Z., & Hartono, Y. (2016). Pengembangan Soal Matematika Model PISA Konten Bilangan untuk Mengetahui Kemampuan Literasi Matematika Siswa. *Jurnal Elemen*, 2(1), 14–26. <https://doi.org/10.29408/jel.v2i1.175>
- Putri, S. S. (2021). *THE DEVELOPMENT OF FOODIVITY INTERACTIVE AS AN INTERACTIVE MULTIMEDIA TO IMPROVE STUDENTS' UNDERSTANDING ON FOOD NUTRITION TOPIC*. Universitas Pendidikan Indonesia.
- Rakhmawati, I. A., & Alifia, N. N. (2018). Kearifan Lokal Dalam Pembelajaran Matematika Sebagai Penguat Karakter Siswa. *Jurnal Elektronik Pembelajaran Matematika*, 5(2), 186–196. <http://jurnal.uns.ac.id/jpm>
- Saputri, H. A., Zuhijrah, Larasati, N. J., & Shaleh. (2023). Analisis Instrumen Assesmen : Validitas, Reliabilitas, Tingkat Kesukaran, dan Daya Beda Butir Soal. *Didaktik : Jurnal Ilmiah PGSD FKIP Universitas Mandiri*, 09(05), 2986–2995.
- Saputri, N. W., Turidho, A., Zulkardi, Z., Darmawijoyo, D., & Somakim, S. (2020). Desain Soal Pisa Konten Uncertainty and Data Konteks Penyebaran Covid-19. *EDU-MAT: Jurnal Pendidikan Matematika*, 8(2), 106–118. <https://doi.org/10.20527/edumat.v8i2.8564>
- Zuhri, N. Z., Syihabuddin, S., & Tatang, T. (2024). Analisis Validitas, Reliabilitas, dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis Artificial Intelligence (AI) melalui Platform QuestionWell. *Jurnal Pendidikan Dan Pembelajaran Indonesia (JPPI)*, 4(2), 693–704. <https://doi.org/10.53299/jppi.v4i2.576>