

Development of Extended Multiple Choice Numeracy Questions Based on The Reeves Model

*Tasya Faricha Amelia¹, Effendi Nawawi², Ratu Ilma Indri Putri³,
Evy Ratna Kartika Waty⁴
^{1, 2, 3, 4} Sriwijaya University
*Email: tasyafa2001@gmail.com

Received: November 2024. **Accepted:** December 2024. **Published:** January 2025.

ABSTRACT

This study aims to develop numeracy questions based on the Reeves development model that produces high-quality numeracy questions. The questions are designed in an expanded multiple-choice format, which allows students not only to choose answers but also to provide in-depth logical reasons. This study uses a design research with a systematic approach involving four main stages in the Reeves model: problem identification and analysis, initial design, prototype development, and implementation and reflection. The research methods include expert validation, small group trials, and field trials involving 26 sixth grade students as research subjects. The results of the study showed that 15 numeracy questions were developed, leaving 10 numeracy questions that were of high quality and had high validity, reliability, varying levels of difficulty, and good discriminatory power. The other 5 numeracy questions were not used because they did not meet the criteria for quality items. The implementation of the questions showed that the expanded multiple-choice format was able to provide an in-depth evaluation of students' abilities and help identify weaknesses in students' understanding in more detail. With these results, this study concludes that the Reeves development model is effective for developing high-quality numeracy questions.

Keywords: Question Development, Numeracy, Expanded Multiple Choice, Reeves Model

How to Cite: Amelia, T., Nawawi, E., Putri, R., & Waty, E. (2025). Development of Extended Multiple Choice Numeracy Questions Based on the Reeves Model. *Journal Of Medives : Journal Of Mathematics Education IKIP Veteran Semarang*, 9(1),

INTRODUCTION

Numeracy skills are an important element in mathematical literacy that every individual, especially students at the elementary school level, must have. Numeracy skills not only include an understanding of numbers and basic mathematical operations, but also the ability to apply these concepts in real-life contexts (Damayanti et al. 2024; Itu, Soro, and Wewe 2024; Sidiq, Ayudia, and Sarjani 2023). In the modern era, numeracy skills play a key role in daily decision making, both in the world of work, education and society (Dilla Nurfadillah et al. 2024; Fajriyah 2022). Therefore, mastering numeracy from an early age is an indicator of educational success.

However, various studies and reports, such as PISA (Programme for International Student Assessment), show that Indonesian students' numeracy literacy skills are still relatively low compared to other countries (D, Khasanah, and Putri 2022; Putrawangsa and Hasanah 2022; Umay, Subali, and Januarsi 2024). This problem is evident in fraction material, especially in the multiplication and division operations of fractions with natural numbers. Students often have difficulty in understanding basic concepts, performing calculations, and connecting this material to the context of everyday life (Fajar Rizqi et al. 2023; Febriyanti and Nurjaman 2023; Widyatma and Ramadhani 2024). For example, many students are unable to answer story-based questions, such as counting ingredients for cooking or dividing items by a certain amount, which involve fraction operations.

One of the causes of low student numeracy skills is the lack of quality questions, both in terms of validity, reliability, level of difficulty, and differentiating power (Hayati and

Nindiasari 2024; Suariantini, Werang, and Astawan 2023). The questions that are often used only focus on memorization and the final answer without exploring the reasons behind the answer (Yunarti and Amanda 2022). This condition shows the need for innovation in developing numeracy questions in order to evaluate students' abilities in more depth.

One innovation that can be done is the use of an expanded multiple-choice format. This format is different from conventional multiple-choice because each answer choice is accompanied by a logical reason. This requires students not only to choose the correct answer but also to understand and explain their choices. With this format, the evaluation can cover the cognitive aspects of students more comprehensively, while also helping to identify the difficulties faced by students in understanding the concept of numeracy.

To produce quality questions, this study adopted the Reeves development model, a systematic development approach that involves problem identification and analysis, initial design, prototype development and implementation and reflection (R, Apriliya, and Kosasih 2018). With this model, it is expected that research can produce valid, reliable numeracy questions, have a proportional level of difficulty, and adequate discriminating power. The ultimate goal is to provide a quality evaluation instrument to support numeracy learning in schools.

METHOD

This study uses a development design or resource design with the Reeves development model. This model involves four systematic stages, namely problem identification and analysis, initial design, prototype development

and implementation and reflection (R et al. 2018). This method is applied to produce numeracy questions in the form of expanded multiple choices that meet the criteria of validity, reliability, level of difficulty and good discriminating power. Data was collected through interviews and test results.

The first stage of the Reeves development model is problem identification and analysis. At this stage, the researcher conducted a needs analysis and literature review so that problems were found regarding numeracy questions that were of poor quality and unable to evaluate students' abilities in depth. So that a solution was obtained to develop expanded multiple-choice questions.

The second stage of Reeves' development model is the initial design. The questions are designed based on contextual discourse text. Each question includes a main question with multiple choice answers and a column for student reasons. So that the assessment is carried out with two approaches, namely in terms of the multiple choice answers and in terms of the answers to the reasons. Multiple choice answers are analyzed using a multiple choice test analysis form where correct answers are worth 1 and wrong answers are worth 0. The answer to the reason is analyzed using a descriptive test analysis form with a 0-5 assessment scale. The following is the reason assessment scale:

Table 1. Scoring Rubric for Reason Answers

Score	Score Rubric
0	There is no answer or a completely wrong answer.
1	There is little effort without using the concepts of multiplication and division of fractions by natural numbers correctly.
2	An attempt to answer is visible, but the concept of multiplying and dividing fractions by natural numbers used is significantly incorrect.
3	Understanding of the concept of multiplying and dividing fractions by natural numbers is evident, but there are some calculation errors.
4	The steps are correct, but there is a small error in the final presentation.
5	All steps and answers are correct.

The third stage of Reeves' development model is prototype development. The items are validated by experts or an expert review is conducted. Validation conducted in an expert review involves assessing content, constructs, and language (Asmara and Sari 2021). After being validated by expert review, the questions are revised and then tested and interviews are conducted on several students or small groups of students to assess the understanding of the questions, the difficulty of the questions, the suitability of the discourse text with everyday life, the

suitability of the images with the discourse text and suggestions.

The fourth stage of Reeves' development model is implementation and reflection. The revised test items based on the results of trials and interviews with a group of students or small groups, were then tested on a larger group of students, namely 26 grade 6 students. The results of the trial that produced two answers, namely multiple choice answers and reason answers, were each analyzed to be assessed in terms of validity, reliability, level of difficulty and discriminatory power.

Validity is calculated using item correlation analysis. The test items are said to be valid if the r table value $> r$ calculated value. The r table used is 3.882. Reliability shows the consistency of the measurement results calculated using the Cornbach Alpha coefficient. The following are the reliability assessment criteria according to Safi'i, Alfi, and Fatih (2024):

Table 2. Reliability Criteria

Reliability Value	Information
< 0.20	Very Low
0.20-0.40	Low
0.41-0.60	Currently
0.61-0.80	Tall
0.81-1.00	Very high

The level of difficulty is indicated by the proportion of students who answer correctly. The following formulas and assessment criteria are used to measure the level of difficulty in multiple-choice answers according to Utami and Adilla (2022):

$$P = \frac{B}{JS}$$

Information :

P =Difficulty level

B =The number of students who answered the questions correctly

JS =Total number of students

Table 3. Multiple Choice Answer Difficulty Level Criteria

Mark P	Information
< 0.30	Difficult
0.31-0.70	Currently
0.71-1.00	Easy

Meanwhile, the formula and assessment criteria used to measure the level of difficulty of the analysis answers according to Astuti, Waluya, and Asikin (2020):

$$TK = \frac{\text{Skor Rata} - \text{Rata Suatu Soal}}{\text{Skor Maksimum Suatu Soal}}$$

Information :
 TK =Difficulty level

Table 4. Criteria Level of Difficulty Answer Reason

Mark TK	Information
< 0.30	Difficult
0.31-0.70	Currently
0.71-1.00	Easy

Discriminatory power measures the extent to which a question can differentiate between students with high and low abilities. The assessment between multiple-choice answers and reason answers to calculate their discriminatory power is the same. The following is the formula and criteria for assessing discriminatory power in multiple-choice answers and analysis answers according to Rahmasari and Ismiyati (2016):

$$DP = \frac{\bar{X}KA - \bar{X}KB}{\text{Skor Maksimal}}$$

Information :

DP =Distinguishing Power

$\bar{X}KA$ =Upper group average

$\bar{X}KB$ =Lower group average

Table 5. Distinguishing Power Criteria

Mark P	Information
< 0.19	Not Good/Discarded
0.20-0.29	Enough
0.30-0.39	Good
$\geq 0,40$	Very good

After being measured in terms of validity, reliability, level of difficulty and discriminating power, it will be determined which questions are of good quality. The criteria for good quality questions are questions that have

content validity according to the curriculum or the questions must be valid, have high reliability (>0.6), have varying levels of difficulty including easy, medium and difficult categories, and adequate discriminating power (>0.2) or the questions have minimal sufficient discriminating power (Sa'idah, Yulistianti, and Megawati 2019).

This research is expected to produce numeracy questions that are not only able to measure students' abilities accurately but also provide a deeper picture of students' way of thinking, especially in understanding and applying numeracy concepts in real life.

RESULTS AND DISCUSSION

The development of multiple-choice numeracy questions was expanded using the Reeves development model consisting of four main stages:

Problem Identification and Analysis

At this stage, the researcher identified and analyzed the problem. Based on the identification and analysis of the problem, it was found that the problem was related to the low-quality numeracy questions and were unable to evaluate students' abilities in depth. So that a solution was obtained to develop expanded multiple-choice questions. The first thing that was done was a needs analysis, the researcher identified the purpose of developing numeracy questions, namely to produce quality numeracy questions to measure students' numeracy abilities validly, reliably, and have a good level of difficulty and differentiating power.

After determining the objectives, the researcher began to conduct a literature review by collecting references on numeracy standards,

expanded multiple-choice question formats and quality question criteria. The researcher also reviewed the concept of validity in terms of construct content and language and also reviewed the concept of reliability, level of difficulty and discriminatory power. The researcher also determined that the indicators of quality numeracy questions to be developed were to have good question criteria in terms of: content validity according to the curriculum, high reliability of more than 0.6, varying levels of difficulty of easy, moderate, difficult and adequate discriminatory power of more than 0.2.

Initial Design

The first thing to do at the design stage is to compile questions. Questions are compiled by determining the competencies being measured such as CP, TP and ATP. Then, the researcher also compiles a question grid containing question indicators and cognitive levels.

After determining the competencies being measured and compiling the question grid. The researcher began to create questions. The compilation of questions was carried out using an expanded multiple-choice format where the answer choices consisted of a combination of options that tested in-depth understanding. Each question included a main question with multiple-choice answers and a column for student reasons. So that the assessment was carried out with two approaches, namely in terms of the multiple-choice answers and in terms of the answers to the reasons. The context used in the questions must also be relevant to the real situation.

The questions consist of 15 expanded multiple-choice questions that have 5 numeracy discourse texts. 1 discourse consists of 3 questions with

different cognitive levels. The following design stage: are the questions developed at the initial

SOAL NUMERASI PILIHAN GANDA DIPERLUAS

Teks Wacana 1 :

Membuat Kue

Ibu membuat kue untuk acara ulang tahun. Ia membutuhkan 4 kg tepung terigu. Setiap adonan kue membutuhkan $\frac{1}{4}$ kg tepung. Berapa banyak adonan kue yang bisa dibuat Ibu?

a. 4
b. 5
c. 6
d. 7

Alasannya :

2. Jika Ibu ingin membuat 10 adonan kue, berapa kg tepung terigu tambahan yang dibutuhkan?

a. 2 kg
b. 3 kg
c. 4 kg
d. 5,5 kg

Alasannya :

7. Berapa banyak baju yang bisa dibuat dari 3 meter kain jika setiap baju membutuhkan $\frac{1}{3}$ meter kain?

a. 4
b. 5
c. 6
d. 7

Alasannya :

8. Jika harga kain batik Rp25.000/meter, berapa uang yang harus dibayarkan untuk membeli kain busbukan agar bisa membuat 5 rok?

a. Rp25.000
b. Rp50.000
c. Rp75.000
d. Rp100.000

Alasannya :

9. Bandingkan kain untuk satu baju dan satu rok. Manakah yang lebih banyak digunakan?

a. Baju 2 × lebih banyak
b. Rok 2 × lebih banyak
c. Sama banyak
d. Tidak cukup data

Alasannya :

15. Jika Budi mencampurkan semua jus yang dibuatnya menjadi satu, berapa banyak gelas jus campuran yang dihasilkan?

a. 3 gelas
b. 4 gelas
c. 5 gelas
d. 6 gelas

Alasannya :

3. Terapa perbandingan banyaknya tepung terigu dan gula pasir yang digunakan untuk satu adonan kue?

a. 1:5
b. 2:5
c. 1:4
d. 4:5

Alasannya :

Teks Wacana 2 :

Membagi Buah

Ibdi memiliki 3 buah semangka. Setiap semangka dipotong menjadi 8 bagian sama besar. Ibdi ingin membagikan potongan semangka kepada teman-temanya.

4. Jika setiap teman mendapat 2 potong semangka, berapa banyak teman yang mendapat bagian?

a. 15
b. 20
c. 25
d. 50

Alasannya :

Teks Wacana 4 :

Membagi Pizza

Andi memiliki 2 buah pizza. Setiap pizza dipotong menjadi 8 bagian sama besar. Andi ingin membagikan pizza tersebut kepada teman-temanya.

10. Jika setiap teman mendapat 3 potong pizza, berapa banyak teman yang mendapat bagian?

a. 4 teman
b. 5 teman
c. 6 teman
d. 7 teman

Alasannya :

11. Jika Andi ingin memberikan $\frac{1}{4}$ bagian dari setiap pizza kepada adiknya, lalu sisanya dibagi rata kepada 4 temannya, berapa bagian pizza yang didapat setiap temannya?

a. 2 bagian
b. 2 $\frac{1}{4}$ bagian
c. 3 bagian
d. 4 bagian

Alasannya :

12. Jika Andi ingin membagikan sisa pizza kepada 3 temannya secara rata, berapa bagian pizza yang akan didapat setiap temannya?

a. 2 bagian
b. 3 bagian
c. 5,2 bagian
d. 5 bagian

5. Jika Ibu ingin membuat jus dari 10 bagian setiap potongan semangka, berapa banyak gelas jus yang bisa dibuat dari 1 buah semangka?

a. 2
b. 4
c. 8
d. 16

Alasannya :

6. Jika adik Ibtis mendapat $\frac{1}{4}$ bagian dari sisa potongan semangka, berapa bagian yang didapat adiknya?

a. 5 bagian
b. 8 bagian
c. 10 bagian
d. 12 bagian

Alasannya :

Teks Wacana 3 :

Membeli Kain

Ibu membeli 5 meter kain untuk membuat baju. Setiap baju membutuhkan $\frac{1}{3}$ meter kain. Ibu juga membeli 2 meter kain untuk membuat rok. Setiap rok membutuhkan $\frac{1}{3}$ meter kain untuk rok.

b. 2,6 bagian
d. 3,6 bagian

Alasannya :

Teks Wacana 5 :

Membuat Jus

Budi memiliki 6 buah jeruk. Setiap buah jeruk menghasilkan $\frac{1}{4}$ gelas jus. Budi juga memiliki 4 buah apel. Setiap buah apel menghasilkan $\frac{1}{4}$ gelas jus.

13. Jika Budi ingin membuat 10 gelas jus jeruk, berapa buah jeruk yang dibutuhkan?

a. 20 buah
b. 32 buah
c. 40 buah
d. 25 buah

Alasannya :

14. Bandingkan banyaknya jus yang dihasilkan dari 1 buah jeruk dan 1 buah apel. Jus mana yang lebih banyak?

a. Jeruk, karena menghasilkan $\frac{1}{4}$ gelas jus
b. Apel, karena menghasilkan $\frac{1}{4}$ gelas jus
c. Sama banyak
d. Tidak cukup data

Prototype Development

At the prototype development stage, this stage focuses on testing questions and improvements based on the results of empirical analysis (Apriatni, Yuhana, and Sukirwan 2022). After conducting an independent

evaluation, the researcher began to consult or validate with an expert review to validate the questions in terms of content, construction, and language. The validators consisted of 3 people who were mathematics teachers who had a master's degree in mathematics education. The following are comments

from the expert review validation and revision:

Table 6. Expert Review Validation Results

Validators	Comment	Revisions and Improvements
Validator 1	Questions need to be supplemented with related discourse text; number 1 is better to go straight to the point; numbers 4 and 10 are too vague.	Add a caption to the discourse text; clarify questions 4 and 10 with wording such as: "How many friends got watermelon/pizza?"
Validator 2	Discourse text 1 needs to be aligned; use consistent terms in number 1; use fraction format in multiple choices; number 8 needs to be clarified.	Refine the wording of discourse text 1; use the consistent term "wheat flour"; use fractions, not decimals, in fraction problems; clarify the logic in problem number 8.
Validator 3	Pay attention to the consistency of the terms in number 1; number 6 is ambiguous; the text of discourse 3 and the image do not match; questions 9 and 12 need to be clarified.	Align the terms "wheat flour" and "flour"; correct number 6 regarding the remaining watermelon pieces; adjust the text of the discourse and the picture; clarify questions number 9 and 12.

After conducting an expert review, the researcher conducted a trial on a group of students or involved a small sample of students to work on the Prototype questions from the results of the revised export validation review. This small sample or small group consisted of 4 grade 6 students. The trial had been conducted, the researcher also conducted interviews with the four students to ask about the students' responses when working on the questions that were developed. The

interview questions included discussions about understanding the questions, the difficulty of the questions, the suitability of the discourse text in the questions with everyday life, the suitability of the images with the discourse text given and suggestions for improvement regarding the questions that were developed. The following are the results of the trial and interviews along with their revisions:

Table 7. Trial Results

Aspect	Student Comments	Revisions and Improvements
Discourse 1: Making Cakes	The measurement from kg to cups is confusing; the image does not support the text.	Convert measurements to cups; add images that match the text to explain ingredient measurements.
Question Number 1	Not straight to the point, hard to understand.	The wording has been clarified to be more direct and easy to understand, while maintaining the same answer choices.
Question Number 2	It's hard to understand.	The question wording has been revised to make it clearer and more relevant to the discourse; the answer choices have not been changed.
Questions	Question number 7 is	The wording of question number 7 has

Number 7 and 8	unclear, number 8 is considered appropriate.	been clarified to make it relevant to the discourse. Number 8 remains intact.
Discourse 5: Making Juice	Not relevant to everyday life; question number 15 is confusing.	The amount of juice from the ingredients has been clarified; the wording of number 15 has been adjusted to make it easier to understand, the answer choices remain the same.
Difficulty Level		Questions that are too difficult and too easy are revised to increase the ideal level of difficulty ($P = 0.3 - 0.7$)
Distinguishing Power		Questions with poor discriminatory power (<2.0) were corrected or discarded.

After the revision, the researcher began to enter the next stage regarding implementation and reflection. The

following are the final results of the test items after expert review validation and small group trials:

Siswa :
Kelas :

Teks Wacana 1 :
Jawablah pertanyaan 1-3 dengan membaca teks wacana 1.

Membuat Kue
Ibu ingin membuat kue untuk acara ulang tahun. Untuk membuat satu adonan kue, ibu biasanya membutuhkan 7 gelas takaran tepung terigu. Satu gelas takaran berisi $\frac{1}{2}$ kg tepung terigu.



1. Jika ibu mempunyai 4 kg tepung, berapa banyak adonan yang ibu buat?
a. 1
b. $\frac{1}{2}$
c. 3
d. $\frac{1}{3}$

Alasannya :

2. Jika ibu ingin membuat 8 adonan kue, berapa kg tepung terigu yang dibutuhkan?
a. 3 kg
b. 6 kg
c. 4 kg
d. 12 kg

Alasannya :

3. Ibu membeli $\frac{1}{2}$ kg tepung terigu di rumah dan ingin membuat 10 adonan kue. Harga tepung terigu adalah Rp14.000/kg. Berapa uang yang harus ibu bayar untuk membeli sisa tepung terigu yang kurang?
a. Rp128.000;
b. Rp130.000;
c. Rp216.000;
d. Rp198.000;

Alasannya :

9. Jika ibu membuat 6 buah roti dan 12 roti. Berapa perbandingan antara kain putih dan kain hitam yang digunakan ibu?
a. Roti 2 : lebih banyak
b. Roti 2 : lebih banyak
c. Sama banyak
d. Tidak cukup data

Alasannya :

Teks Wacana 4 :
Jawablah pertanyaan 10-12 dengan membaca teks wacana 4.

Membagi Pizza



Andi memiliki 2 buah pizza. Setiap pizza dipotong menjadi 8 bagian sama besar. Andi ingin membagikan pizza tersebut kepada kuzin-kuzinnya.

10. Jika setiap kuzin mendapat 4 potong pizza, berapa banyak kuzin Andi yang mendapatkan pizza?
a. 4 kuzin
b. 5 kuzin
c. 6 kuzin
d. 7 kuzin

Alasannya :

Teks Wacana 2 :
Jawablah pertanyaan 4-6 dengan membaca teks wacana 2.

Membagi Buah



Budi memiliki 5 buah semangka. Setiap semangka dipotong menjadi 8 bagian sama besar. Budi ingin membagikan potongan semangka kepada teman-temannya.

4. Jika setiap teman mendapat 2 potong semangka, berapa banyak teman Budi yang mendapatkan semangka?
a. 15
b. 20
c. 25
d. 30

Alasannya :

5. Jika Budi ingin membuat jus dari $\frac{1}{2}$ bagian setiap potongan semangka, berapa banyak gelas jus yang bisa dibuat dari 1 buah semangka?
a. 2
b. 4
c. 8
d. 16

Alasannya :

11. Jika Andi ingin memberikan $\frac{1}{2}$ bagian dari setiap pizza kepada adiknya, lalu sisa pizzanya dibagi rata kepada 4 temannya, berapa potong pizza yang didapat setiap temannya?
a. 2 bagian
b. $2\frac{1}{2}$ bagian
c. 3 bagian
d. 4 bagian

Alasannya :

12. Jika Andi selanjutnya memakan $\frac{1}{2}$ dari pizza yang dia punya. Kemudian sisanya ia bagikan kepada 4 temannya. Berapa potong pizza yang diperoleh temannya Andi?
a. 2 bagian
b. $\frac{1}{2}$ bagian
c. 1 bagian
d. $\frac{1}{2}$ bagian

Alasannya :

Teks Wacana 5 :
Jawablah pertanyaan 13-15 dengan membaca teks wacana 5.

Membuat Jus



Budi memiliki 12 buah jeruk. Satu buah jeruk menghasilkan $\frac{1}{2}$ gelas jus. Budi juga memiliki 16 buah apel. Satu buah apel menghasilkan $\frac{1}{4}$ gelas jus.

13. Jika Budi ingin membuat 9 gelas jus jeruk, berapa buah jeruk yang dibutuhkan?
a. 10 buah
b. 9 buah
c. 12 buah
d. 4 buah

Alasannya :

6. Jika adik Budi mendapat $\frac{1}{2}$ bagian dari semua potongan semangka, berapa bagian yang didapat adiknya Budi?
a. 50 bagian
b. 30 bagian
c. 10 bagian
d. 25 bagian

Alasannya :

Teks Wacana 3 :
Jawablah pertanyaan 7-8 dengan membaca teks wacana 3.

Membeli Kain



Ibu membeli 3 meter kain polos dan 2 meter kain batik yang masing-masing akan digunakan untuk membuat baju dan rok. Untuk membuat satu baju dibutuhkan $\frac{1}{2}$ meter kain polos. Untuk membuat satu rok dibutuhkan $\frac{1}{2}$ meter kain batik.

7. Berapa banyak baju yang bisa dibuat dari bahan yang dimiliki ibu sekarang?
a. 4
b. 5
c. 6
d. 7

Alasannya :

8. Jika harga kain batik Rp25.000/meter, berapa uang yang harus dibayarkan untuk membeli kain batik tambahan agar bisa membuat 16 rok?
a. Rp15.000
b. Rp50.000
c. Rp75.000
d. Rp100.000

14. Bandingkan banyaknya jus yang dihasilkan dari 1 buah jeruk dan 1 buah apel. Jus mana yang lebih banyak?
a. Jeruk, karena menghasilkan $\frac{1}{2}$ gelas jus
b. Apel, karena menghasilkan $\frac{1}{4}$ gelas jus
c. Sama banyak
d. Tidak cukup data

Alasannya :

15. Jika Budi membuat jus dengan mencampurkan semua bahan yaitu jeruk dan apel yang Budi punya menjadi satu, berapa banyak gelas jus campuran yang dihasilkan?
a. 13 gelas
b. 24 gelas
c. 15 gelas
d. 9 gelas

Alasannya :

Implementation and Reflection

At this final stage, the researcher's focus is on the application of questions and a definite assessment of whether the questions developed are valid, reliable, have a good level of difficulty and differentiating power or not (Fidia, Puspitawati, and Yakub 2022). The application of the questions was carried out in class 6 at SD Negeri 19 Talang Kelapa which consisted of 26 students. During the application, it was seen that there were students who found it easy to do the numeracy questions and there were students who found it difficult to do the numeracy questions.

After the application or implementation of the questions to students, the results of the application are analyzed to ensure whether the questions are valid and reliable and have a good level of difficulty and differentiating power. Because the test

used is an expanded multiple choice where not only multiple choices are assessed and analyzed, but there are logical reasons that must also be assessed and analyzed. So the analysis is carried out using two methods, namely analysis of the form of multiple choice questions and analysis of the answer reasons using the analysis of the essay test form. Both analyzes are carried out by analyzing the validity, reliability of the level of difficulty, and differentiating power. The following are the results of the analysis:

1) Validity

The results of the study on the validity of the questions in the Multiple Choice Analysis showed that 10 questions were declared valid and 5 questions were declared invalid. However, the Reason analysis showed that 15 questions developed were valid. The distribution of 15 questions based on the two validity analyses is as follows:

Table 8. Item Distribution Based on Item Validity

Types of Analysis	Validity Index	Question Items	Amount	Presentation
Multiple Choice Analysis	Rcount ≥ 0.338 (valid question)	2,5,7,9,1 0,11,12,13,14,15	10	66.67%
	Rcount < 0.338 (invalid question)	1,3,4,6,8	5	33.3%
	Rcount ≥ 0.338 (valid question)	1,2,3,4,5,6,7,8,9,10,1 1,12,13,14,15	15	100%
Reason Analysis	Rcount < 0.338 (invalid question)			
	0	0	0%	
		Multiple Choice Analysis	0.66	Tall
		Reason Analysis	0.88	Very high

2) Reliability

The following are the results of the reliability values in both analyses:

Table 9. Item Distribution Based on Reliability

Types of Analysis	Reliability	Reliability Level
-------------------	-------------	-------------------

3) Difficulty Level

The following is the distribution of 15 questions based on their level of difficulty in both analyses:

Table 10. Distribution of Items Based on Difficulty Level

Type	Diffic	Question	Amo	Present
------	--------	----------	-----	---------

s of Anal ysis	ulty Index	Items	unt	ation
Multi ple	Easy	4,6,7,10,13,14	6	40%
Choi ce	Curre ntly	1,2,3,5,11,15	6	40%
Anal ysis	Diffic ult	8,9,12	3	20%
Reas on	Easy	4,5,6,10,14	5	33.3%
Anal ysis	Curre ntly	1,2,3,7,11,12,13,15	8	53.3%
	Diffic ult	8.9	2	13.3%

4) Distinguishing Power

The following are the results of the discriminatory power analysis based on multiple choice and reasons:

Distribution of Essay Question Items Based on Distinguishing Power

Table11. Distribution of Grains Based on Discriminating Power

Types of Analysis	Distinguishing Power Index	Question Items	Amount	Presentation
Multiple Choice Analysis	Not good	3,4,5,6,8	4	26.6%
	Enough	7,14,15	3	20%
	Good	1,9,10,12,13	5	33.3%
	Very good	2,5,11	3	20%
Reason Analysis	Not good	1,3,8	3	20%
	Enough	2,4,9,14	4	26.6%
	Good	6,7,11,12,13,15	6	40%
	Very good	5.10	2	13.3%

Based on the four analyses, the following is a combination of validity,

reliability, level of difficulty and differentiating power analyses:

Table 12. Combined Analysis

No. Questio n	Validity		Reliability		Difficulty Level		Distinguishing Power	
	Multipl e Choice Analysi s	Reason Analysi s	Multipl e Choice Analysi s	Reason Analysi s	Multiple Choice Analysis	Reason Analysis	Multipl e Choice Analysi s	Reason Analysi s
1	Invalid	Valid		Very high	Currentl y	Currentl y	Good	Not good
2	Valid	Valid	Tall	Very high	Currentl y	Currentl y	Very good	Enough
3	Invalid	Valid		Very high	Currentl y	Currentl y	Not good	Not good
4	Invalid	Valid		Very high	Easy	Easy	Not good	Enough
5	Valid	Valid	Tall	Very	Currentl	Easy	Very	Very

				high	y		good	good
6	Invalid	Valid		Very high	Easy	Easy	Not good	Good
7	Valid	Valid	Tall	Very high	Easy	Currentl y	Enough	Good
8	Invalid	Valid		Very high	Difficult	Difficult	Not good	Not good
9	Valid	Valid	Tall	Very high	Difficult	Difficult	Good	Enough
10	Valid	Valid	Tall	Very high	Easy	Easy	Good	Very good
11	Valid	Valid	Tall	Very high	Currentl y	Currentl y	Very good	Good
12	Valid	Valid	Tall	Very high	Difficult	Currentl y	Good	Good
13	Valid	Valid	Tall	Very high	Easy	Currentl y	Good	Good
14	Valid	Valid	Tall	Very high	Easy	Easy	Enough	Enough
15	Valid	Valid	Tall	Very high	Currentl y	Currentl y	Enough	Good

Quality numeracy questions are questions that have content validity according to the curriculum or the questions must be valid, high reliability (>0.6), varying levels of difficulty including easy, medium, and difficult categories, and adequate discriminatory power (>0.2) or the questions have minimal sufficient discriminatory power. This study shows the importance of ensuring the quality of numeracy questions through analysis of validity, reliability, level of difficulty, and discriminatory power. Based on the combined analysis table, it can be seen that out of 15 questions developed, 10 questions meet the criteria as quality questions based on content validity according to the curriculum, high reliability, distribution of varying levels of difficulty, and adequate discriminatory power. The 10 questions are numbers 2, 5, 7, 9, 10, 11, 12, 13, 14, and 15. The other 5 numeracy questions were not used because they did not meet the criteria for quality items.

Content validity is an important

element because it ensures that each question reflects the competency measured according to the curriculum. In this study, validity was strengthened through multiple choice and reason analysis, which provided a deeper measurement dimension. Of the 15 questions developed, only 10 were considered valid based on the multiple choice analysis. However, in the reason analysis, all 15 questions were declared valid. This shows that the question format that includes reasons can provide an additional, more comprehensive measurement dimension. Content validity is also often reviewed through expert review to ensure its relevance to curriculum standards and the abilities being measured. A study by Damayanti, Daryono, and Hari Rayanto (2023) states that content validity is an aspect that must be evaluated by experts in the relevant field so that questions can be ensured to be relevant to learning objectives.

In addition, high item reliability indicates instrument consistency, so that the measurement results remain accurate even when used repeatedly. In

this study, item reliability reached a high value (0.66 for multiple choice) to very high (0.88 for reasons). This value indicates that the item instrument can be used repeatedly with consistent results under similar conditions. Previous studies, such as those conducted by Kania et al. (2023), revealed that high reliability is essential to ensure the accuracy of evaluation. This is especially true in the context of primary education, where evaluation instruments must be able to capture student performance consistently.

The distribution of difficulty levels is also a concern, with questions spanning easy, medium, and hard categories. This ensures that the questions can effectively evaluate students with different levels of ability. Questions that are too easy or too hard are avoided to maintain balance and relevance of the evaluation. An even distribution ensures that the questions can evaluate students with different levels of ability. For example, questions that are too easy tend not to be able to distinguish students with high ability from those with low ability, while questions that are too hard can hinder student motivation. This is in line with the view Utami and Adilla (2022), which states that an ideal question should have a difficulty index between 0.3 and 0.7 to ensure diversity.

Discriminating power is also a key indicator, where questions with low discriminating power values are improved or removed so that the instrument can differentiate students with diverse understandings. In this study, questions with low discriminating power were improved or removed to maintain the quality of the instrument. Sufficient discriminating power (>0.2) ensures that the evaluation instrument can identify various levels of student ability. According to Susilowati

et al. (2024), discriminatory power is a critical aspect in item development because it provides insight into the effectiveness of the item in assessment.

This research is in line with various studies that emphasize the importance of item analysis. For example, Magdalena et al. (2021) showed that validity, reliability, level of difficulty, and discriminatory power are key parameters to ensure that test questions are effective as evaluation tools. This is supported by the results of the analysis which highlighted that 10 out of 15 questions met the criteria as high-quality questions, with adequate levels of validity, reliability, difficulty, and discriminatory power. Another study by Cantika et al. (2024) also found that evaluation instruments that do not meet any of these criteria can reduce the quality of evaluation results. Therefore, it is important to ensure that the questions are not only in accordance with the curriculum but also able to provide meaningful information about student abilities.

CONCLUSION

This study successfully developed numeracy questions based on the Reeves development model with an expanded multiple-choice format. The results of the study revealed that out of 15 numeracy questions developed, only 10 questions met the criteria for good quality, namely having high validity and reliability, varying levels of difficulty, and adequate discriminatory power. This expanded multiple-choice format allows students not only to choose the correct answer, but also to provide in-depth logical reasons, providing a comprehensive evaluation of students' understanding and numeracy abilities. The Reeves development model has proven

effective in producing valid, reliable, and contextual numeracy questions. This study provides significant contributions for teachers and researchers in designing more quality and in-depth numeracy learning evaluation tools.

REFERENCE

- Apriatni, Sri, Yuyu Yuhana, and Sukirwan Sukirwan. 2022. "Pengembangan Instrumen Literasi Numerasi Materi Trigonometri Kelas X Sma." *EDU-MAT: Jurnal Pendidikan Matematika* 10(2):185. doi: 10.20527/edumat.v10i2.13720.
- Asmara, Adi, and Debby Juita Sari. 2021. "Pengembangan Soal Aritmetika Sosial Berbasis Literasi Matematis Siswa SMP." *Jurnal Cendekia : Jurnal Pendidikan Matematika* 5(3):2950–61. doi: 10.31004/cendekia.v5i3.982.
- Astuti, Astuti, Stevanus Budi Waluya, and Mohammad Budi Asikin. 2020. "Instrumen Kemampuan Berpikir Kreatif Matematika Untuk Siswa Kelas IV Sekolah Dasar." *Musamus Journal of Primary Education* 3(1):27–34. doi: 10.35724/musjpe.v3i1.3117.
- Cantika, Putu Maha Iswari Putri, Luqman Hakim, Vivi Pratiwi, and Citra Auliyatun Nazwah. 2024. "PELAJARAN EKONOMI BISNIS DAN ADMINISTRASI UMUM KELAS XI SMK MENGGUNAKAN PROGRAM ANATES VERSI 4 . 0." *Jurnal Ilmiah Kajian Multidisipliner* 8(11):324–41.
- D, Darwanto, Mar'atun Khasanah, and Anggi Monica Putri. 2022. "Penguatan Literasi, Numerasi, Dan Adaptasi Teknologi Pada Pembelajaran Di Sekolah." *Eksponen* 11(2):25–35. doi: 10.47637/eksponen.v11i2.381.
- Damayanti, Ayu Maya, Daryono Daryono, and Yudi Hari Rayanto. 2023. *Evaluasi Pembelajaran*. Cetakan Pe. edited by E. Tresna Setiawan. CV. Basya Media Utama.
- Damayanti, Elvina, Dinda Zastia Alviony, Aprilia Dwi Putri, Rita Feni, Program Studi, Pendidikan Matematika, Program Studi, Pendidikan Ekonomi, and Program Studi. 2024. "Sosialisasi Dan Pelatihan Metode Jarimatika Untuk Meningkatkan Kemampuan Numerasi Siswa Sekolah Dasar." 4(5):490–95.
- Dilla Nurfadillah, Firyal Nasywa Aufa, and Ichsan Fauzi Rachman. 2024. "Membangun Kualitas Pendidikan Melalui Kemampuan Literasi Dan Numerisasi Dalam Implementasi Kurikulum Merdeka." *ALFIHRIS : Jurnal Inspirasi Pendidikan* 2(3):128–40. doi: 10.59246/alfihris.v2i3.876.
- Fajar Rizqi, Ardhian, Bilqis Luthfi Adilla, Erani Sulistiyawati, and Taufiqurrohman. 2023. "Analisis Kesulitan Belajar Matematika Pada Siswa Sekolah Dasar Dan Alternatif Pemecahannya." *Jurnal Pendidikan Dasar Flobamorata* 4(1):481–88. doi: 10.51494/jpdf.v4i1.588.
- Fajriyah, Euis. 2022. "Kemampuan Literasi Numerasi Siswa Pada Pembelajaran Matematika Di Abad 21." *Seminar Nasional Pendidikan* 21:403–9.
- Febriyanti, Natasya, and Asep Rudi Nurjaman. 2023. "Analisis Kesalahan Siswa Dalam Menyelesaikan Soal Cerita Matematika Di Sekolah Dasar." <https://ejournal.iaifa.ac.id/index.php/Dirasah> Accepted: 6(2):322–28.

- Fidia, Farah, Rinie Pratiwi Puspitawati, and Pramita Yakub. 2022. "Pengembangan Instrumen Soal Higher Order Thinking Skills (HOTS) Materi Jaringan Dan Organ Pada Tumbuhan Kelas XI SMA." *Berkala Ilmiah Pendidikan Biologi (BioEdu)* 11(3):745–54. doi: 10.26740/bioedu.v11n3.p745-754.
- Hayati, Rohimatul, and Heps Nindiasari. 2024. "Pengembangan Instrumen Asesmen Kompetensi Minimum Numerasi Pada Domain Data Dan Ketidakpastian Untuk Siswa SMP." 6(2):84–93.
- Itu, Maria Alexandria, Viorentina Meo Soro, and Melkior Wewe. 2024. "Profil Kemampuan Numerasi Siswa Sekolah Dasar Di SDK Kisanata." *Polinomial : Jurnal Pendidikan Matematika* 3(2):107–13. doi: 10.56916/jp.v3i2.921.
- Kania, Nia, Rinovian Rais, Yance Manoppo, Zuli Nuraeni, Ahmad Ahmad, Abdul Munif, Yeti Sulfiati, Anas Anas, Mursidin Mursidin, Reni Suwenti, Geubrina Maghfirah, Hanida Listiani, and Dedi S. 2023. *Evaluasi Pendidikan (Sebuah Tinjauan Kritis)*. CV. Edupedia Publisher.
- Magdalena, Ina, Septy Nurul Fauziah, Siti Nur Faziah, and Fika Sulaehatun Nupus. 2021. "Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas Iii Sdn Karet 1 Sepatan." *BINTANG : Jurnal Pendidikan Dan Sains* 3(2):198–214.
- Putrawangsa, Susilahudin, and Uswatun Hasanah. 2022. "Analisis Capaian Siswa Indonesia Pada PISA Dan Urgensi Kurikulum Berorientasi Literasi Dan Numerasi." *Jurnal Studi Pendidikan Dan Pembelajaran* 1(1):1–12.
- R, Wili Karlina, Seni Aprilia, and E. Kosasih. 2018. "Pengembangan Soal Pada Pembelajaran Penulisan Huruf Kapital Dalam Teks Cerita Pendek." *Jurnal Ilmiah Pendidikan Guru Sekolah Dasar* 5(4):169–77.
- Rahmasari, Dias, and Ismiyati. 2016. "Analisis Butir Soal Mata Pelajaran Pengantar Administrasi Perkantoran." *Economic Education Analysis Journal* 5(1):317–30.
- Sa'idah, Nusrotus, Hayu Dian Yulistianti, and Eka Megawati. 2019. "Analisis Instrumen Tes Higher Order Thinking Matematika Smp." *Jurnal Pendidikan Matematika* 13(1):41–54. doi: 10.22342/jpm.13.1.6619.41-54.
- Safi'i, Mochammad Dwi Irvan, Cindya Alfi, and Mohamad Fatih. 2024. "Jurnal Perseda." *Jurnal Persada VII*(2):195–205.
- Sidiq, Fadhil, Inge Ayudia, and Tri Mustika Sarjani. 2023. "Optimalisasi Gerakan Literasi Sekolah Melalui Desain Kelas Literasi Numerasi Di Sekolah Dasar Kota Langsa." *Journal of Human and Education* 3(3):69–75.
- Suariantini, N. N. G., B. R. Werang, and I. G. Astawan. 2023. "Instrumen Asesmen Numerasi Online Menggunakan Aplikasi Kahoot Pada Mata Pelajaran Matematika Kelas IV Sekolah Dasar." *Innovative: Journal Of ...* 3(2):5712–5824.
- Susilowati, Dwi Andrianik, Nidaria Saputri, Khafidhotun Ni'mah, Luqman Hakim, and Vivi Pratiwi. 2024. "PENERAPAN APLIKASI ANATES UNTUK MENGANALISIS KUALITAS SOAL HOTS PADA FASE-E ELEMEN PENGGUNAAN APLIKASI PENGOLAH ANGKA (SPREADSHEET) DALAM

EVALUASI PEMBELAJARAN :
STUDI KASUS PADA SISWA
KELAS X AKUNTANSI DI SMK
NEGERI 1 KENDAL.” 8(11):272–
83.

Umaya, F., B. Subali, and T. D. Januarsi. 2024. “Peningkatan Literasi Numerasi Siswa Materi Usaha, Energi Dan Pesawat Sederhana Melalui Model Pembelajaran Discovery Learning Pada Kelas VIII SMPN 16” ... *Nasional Pendidikan Dan ...* 649–57.

Utami, Lisa, and Raysha Adilla. 2022. “ANALISIS KETERAMPILAN PROSES SAINS SISWA MENGGUNAKAN VIRTUAL LABORATORY PHYSICS EDUCATION TECHNOLOGY (PhET) PADA MATERI INDIKATOR ASAM BASA.” *Journal of Research and Education Chemistry* 4(1):50. doi: 10.25299/jrec.2022.vol4(1).9348.

Widyatma, Yollanda Vannesicha, and Amanda Diva Hadi Ramadhani. 2024. “Analisis Kemampuan Pemecahan Masalah Matematis Pada Materi Bilangan Dan Aljabar Siswa Kelas IV SDN 4 Piji.” *Jurnal Pendidikan Dan Pembelajaran* 3(01):335–49.

Yunarti, Tina, and Ari Amanda. 2022. “Pentingnya Kemampuan Numerasi Bagi Siswa.” *Seminar Nasional Pembelajaran Matematika, Sains Dan Teknologi* 2(1):44–48.