

Development of Numeracy Problems Based on Education for Sustainable Development (ESD) to Measure Critical Thinking Ability

Dian Kurniati ¹, Intan Wahyuna ¹, Dhanar Dwi ¹, Percy Sepeng ², Sharifah Osman ³

¹ University of Jember, Jl. Kalimantan No 37, Jember, Indonesia

² Sol Plaatje University, 10 Jan Smuts Blvd, Kimberley, South Africa

³ Universiti Teknologi Malaysia, Jalan Iman, 81310 Skudai, Malaysia

Abstract – One of the 21st-century skills that students must possess is critical thinking. An instrument is needed to assess the level of success in developing students' critical thinking abilities. The contribution of this research is the development of numeracy problems based on Education for Sustainable Development (ESD) that can measure students' critical thinking abilities. This research is quantitative, with analysis using Ministep software and involving 75 students as research subjects. Numeracy problems were developed using the 3-D model with the stages of define, design, and develop. The results of this study indicate that the 15 developed numeracy problems are valid as they meet at least one criterion of validity and reliability, with an item reliability value of 0.96. This suggests that the instrument can effectively measure critical thinking skills, cover a diverse range of difficulty levels, and ensure that the formulated questions are suitable for measuring and assessing various student abilities. These results indicate that the ESD-based numeracy problems developed can be used as reference material in similar research or studies related to the analysis of students' critical thinking abilities in solving ESD-based numeracy problems.

Keywords – Question quality, ESD-based numeracy, critical thinking abilities, mathematics education.

1. Introduction

21st-century skills have become an integral part of the vocabulary in the field of education, considering the changes and technological advancements that require individuals to possess complex competencies and abilities. One of the seven life skills needed in the 21st century is critical thinking and problem-solving skills [1]. Other perspectives also mention that 21st-century skills encompass various soft skills and dispositions, including cross-cultural skills, collaboration skills, critical thinking, and problem-solving skills [2]. These perspectives highlight critical thinking as a necessity for everyone living in the 21st century. Critical thinking is reasoned reflective thinking focused on determining what to believe or do [3].

Critical thinking skills have been recognized as one of the most important thinking skills and a crucial indicator of the quality of students' learning [4]. The field of knowledge capable of developing critical thinking, creativity, collaboration, and communication skills is mathematics education. However, the importance of critical thinking skills is not aligned with the mathematics proficiency of Indonesian students, which is still below international standards. The results of the OECD's 2022 PISA study indicate that the average score of Indonesian students in mathematics proficiency is 366, compared to the OECD's average score of 472 [5]. These results are not significantly different from the PISA 2018 results, which showed that Indonesia is still in the low-performance quadrant with high equity. Therefore, there is still an opportunity for Indonesia to improve critical thinking skills as it possesses untapped capacity and potential [6].

DOI: 10.18421/TEM133-19

<https://doi.org/10.18421/TEM133-19>

Corresponding author: Dian Kurniati,
University of Jember, Jl. Kalimantan No 37,
Jember, Indonesia


Email: dian.kurniati@unej.ac.id

Received: 06 February 2024.

Revised: 07 May 2024.

Accepted: 24 May 2024.

Published: 27 August 2024.

 © 2024 Dian Kurniati et al; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDeriv 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

The cognitive ability of students in critical thinking can be measured and observed through the results of formative tests [6]. The importance of measuring critical thinking abilities lies in the fact that critical thinking is an essential skill that can serve as an indicator of success in achieving competency standards in the learning process [8]. Measuring critical thinking skills not only functions as an indicator of success in education but also serves as a foundation for continuous improvement in education, leading to an enhancement in the quality of thinking and the preparedness of students to face complex changes in the future.

Appropriate and effective assessment tools are necessary for measuring students' critical thinking abilities. These assessment tools must be well-designed to comprehensively identify students' critical thinking abilities, referring to key indicators of critical thinking. The core activities of critical thinking include analysis, evaluation, and further argumentation [9]. One assessment related to critical thinking abilities is the Minimum Competency Assessment (AKM) [10]. The minimum competency assessment involves numeracy as one of its components. Numeracy is the ability to think using mathematical concepts, procedures, facts, and tools to solve everyday problems in various relevant contexts for individuals as citizens of Indonesia and the world [11].

Abstract numeracy skills within the context of mathematics learning are a prerequisite for leveraging mathematical thinking abilities in real-life situations [12]. Numeracy questions used in assessments can encompass various contexts and demand critical thinking relevant to real-world challenges. Therefore, students are not only assessed on their mathematical abilities but also on their ability to apply critical thinking in complex and diverse situations. Students with high numeracy skills demonstrate better critical thinking abilities in completing critical thinking tests compared to students with low numeracy skills [13]. Hence, to measure the extent of students' critical thinking abilities, it is essential to examine their skills in solving numeracy problems.

The ability to interpret problems in real life is closely related to numeracy questions. In this context, students are not only exposed to understanding mathematical concepts but also to the real-life application of numeracy in everyday situations. These aspects include knowledge, skills, attitudes, and values that are highly relevant to shaping a sustainable future.

Education for Sustainable Development (ESD) has become a key element in supporting the sustainable development goals (SDGs) program.

ESD aims to provide learners with knowledge, skills, values, and the ability to make informed decisions and act responsibly for environmental integrity, economic sustainability, and a fair society for the present and future generations, respecting cultural diversity [14]. This makes questions with ESD concepts require learners to think critically as they involve analysis, evaluation, and problem-solving in complex situations relevant to current global issues and those that may arise in the future.

Within the context of numeracy questions based on education for sustainable development (ESD), learners are confronted with more complex critical thinking tasks. They are not only required to understand and apply mathematical concepts but also to analyze, evaluate, and solve problems in complex situations relevant to current global issues and those that may arise in the future. Thus, the development of numeracy questions integrating ESD concepts not only enhances mathematical literacy but also stimulates the development of students' critical thinking skills.

One of the reasons for the low mathematics proficiency of Indonesian students, as mentioned earlier in the PISA 2022 results, is the lack of students' exposure to numeracy-based exercise questions designed to improve their skills [15]. Insufficient exposure to this type of question can have a negative impact on the development of students' numeracy literacy, coupled with a lack of practice and exposure to mathematical contexts in everyday life.

Teachers' skills in designing and presenting numeracy-based questions to train students' numeracy literacy are also crucial [16]. Therefore, improving teachers' skills in creating relevant and motivating numeracy exercises can play a vital role in enhancing students' math achievements. Despite numerous studies on the development of evaluation instruments and math exercises, there are limitations in research focusing on the development of numeracy questions based on education for sustainable development (ESD). In facing the complexity of current global challenges, integrating ESD concepts into numeracy questions becomes increasingly important to provide a holistic and relevant learning approach. Therefore, further research on the development of numeracy questions based on ESD is expected to enrich the literature and make a significant contribution to improving the mathematics skills of Indonesian students.

The development of numeracy questions based on ESD remains interesting for expansion, given the need for sustainable development and the rapidly changing landscape that demands instruments relevant to the skills to be measured.

Several studies on the development of ESD-based numeracy questions in Indonesia mostly focus on elementary school students; hence, there is a need for development at the secondary school level. Therefore, this research aims to generate ESD-based numeracy questions for Grade XI high school students.

The main objective of this research is to produce ESD-based numeracy questions that can be used as a tool to measure the critical thinking abilities of high school students. The developed questions are expected to meet high validity and reliability criteria. Thus, the research questions in this study are: 1) What is the process of developing ESD-based numeracy questions to measure critical thinking abilities? 2) What are the results of developing ESD-based numeracy questions to measure critical thinking abilities that can be considered valid and reliable?

Through this research, it is anticipated to make a positive contribution to the development of evaluation instruments relevant to the needs of sustainable development, especially at the high school level, and serve as a foundation for the development of numeracy literacy and critical thinking skills among students in Indonesia.

2. Methodology

The type of research used is research and development (R&D). Numeracy questions based on education for sustainable development are designed and developed using the "3-D Model," which stands for its four main stages: define, design, and develop [17]. The first stage is define, aiming to identify and formulate learning requirements and needs. There are five steps in this stage, including (1) front-end analysis, (2) student analysis, (3) task analysis, (4) concept analysis, and (5) objective specification.

The next stage is design, with the goal of designing a prototype of numeracy questions to measure critical thinking abilities. There are four steps in this stage, including (1) test blueprinting, (2) media selection, (3) format selection, and (4) initial design.

The process then proceeds to development. Components created in the previous stages need to be modified before becoming the final version. This stage aims to gather feedback through formative evaluation, which is then revised. There are two main steps in this stage, including (1) expert validation, which assesses the feasibility of the prototype product by competent validators in their field, and (2) limited testing and field testing. Limited testing involves testing the product on a small scale to identify imperfect parts for revision based on question readability and student feedback.

Testing and revisions are repeated until the product is consistent. In this study, limited testing was conducted with 7 students, while field testing aimed to analyze quality criteria, namely validity and reliability using Rasch modelling. Field testing in this study involved 75 students from Grade XI at Senior High School in Bondowoso, Indonesia.

The instruments, aspects assessed, and respondents in the study can be seen in Table 1.

Table 1. Instruments, aspects assessed, and respondents

Instrument	Aspects Assessed	Respondents
Prototype of numeracy questions based on ESD	Alignment of numeracy questions with the AKM 2021 framework, ESD, and critical thinking ability indicators	Conducted by the researcher
Expert validation sheet	Validity of numeracy question devices	Expert validators
Question readability questionnaire	Readability of questions	Limited test subjects
Prototype of 3 numeracy questions based on ESD	Valid and reliable	Field test subjects
Instrument	Aspects Assessed	Respondents

The data collection method in this research involves the use of questionnaires and tests. The questionnaire method includes a validation sheet for experts and a questionnaire on question readability through Google Forms during the limited test. The validity level of the data resulting from expert validation of the prototype of numeracy questions based on ESD is assessed using the validity assessment steps [18]. Meanwhile, the test method is conducted during the limited and field tests. The results of the field test are then analyzed using Rasch analysis.

The quantitative validation process includes a review of the following aspects: (1) mean square outfit (MNSQ) values accepted: $0.5 < MNSQ < 1.5$, (2) standardized z-values (ZSTDQ) accepted: $-2.0 < ZSTD < 2.0$, (3) point measure correlation (Pt Measure Corr) values accepted: $0.4 < Pt Measure < 0.85$ [19], [20]. Questions that are considered valid must meet at least one of these criteria [20].

Reliability analysis is conducted by considering Cronbach's alpha value, person reliability, and item reliability. Cronbach's alpha value (n) is used to measure the overall reliability of the interaction between numeracy questions and respondents. If the value of n is less than 0.5, it can be considered poor. If the value is in the range of 0.5 to less than 0.6, it can be categorized as poor.

If the value is in the range of 0.6 to less than 0.7, reliability is considered sufficient. If the value is in the range of 0.7 to less than 0.8, it can be categorized as good. Meanwhile, if the value of n reaches 0.8 or more, reliability is considered excellent [21].

The person reliability (P) value is used to determine the consistency of respondent answers, while the item reliability (P) value is used to assess the quality of items in numeracy questions. If the P value is less than 0.67, reliability is considered weak. If the P value is in the range of 0.67 to less than 0.80, reliability can be categorized as sufficient. If the P value is in the range of 0.81 to less than 0.90, reliability is considered good. If the P value is in the range of 0.91 to less than 0.94, reliability is categorized as excellent. Finally, if the P value is 0.94 or higher, reliability is considered outstanding [21].

The Rasch model involves only one logistic parameter, which is the difficulty level of the questions. From the field test data, the difficulty level of each question can be analyzed. The difficulty level of each question is indicated on the item map in the form of a vertical graph. Measure value less than -1 indicates a very easy item, a value between -1 and 0 is categorized as an easy item, a value between 0 and 1 is categorized as a difficult item, and a measure value above 1 indicates a very difficult item [21]. In the Ministep software, person measure indicates the average value of respondents in the instrument. An average value greater than logit 0.0 means that respondents tend to answer correctly on various numeracy questions.

Student score data from the field test results are analyzed to group students through person grouping. Person grouping can be identified through the separation value. The higher the separation value, the better the quality of the instrument in terms of overall respondents and numeracy questions. To separate individual groups, the equation $H = [4 \times \text{SEPARATION} + 1] : 3$ can be used.

3. Results

Numeracy questions based on education for sustainable development to measure critical thinking skills are developed using the "4-D Model" with modifications [17]. In this research, the development process consists of the definition stage, the design stage, and the development stage. The first stage is the definition stage, starting with front-end analysis. In this stage, a preliminary study is conducted on the conditions of the 21st century, which require individuals to have critical thinking and problem-solving skills.

Critical thinking skills are important to measure as they serve as a foundation for continuous improvement in education.

Numeracy questions are considered a suitable tool for measuring critical thinking skills. These numeracy questions are integrated with education for sustainable development because questions with this concept will require learners to think critically, involving analysis, evaluation, and problem-solving in complex situations relevant to global issues. Additionally, the current development of numeracy questions is limited, necessitating the creation of new numeracy question products.

The next step is student analysis. The results of the PISA study released by the OECD in 2022 indicate that the average score of Indonesian students in mathematics is not significantly different from previous PISA results, which were below average and categorized as low performance with high equity. This suggests that there is still an opportunity for Indonesia to improve critical thinking skills. The numeracy questions used in the national assessment in Indonesia follow the AKM model; therefore, the developed numeracy questions will also use the AKM model. In the task analysis stage, the measurement to be conducted focuses on critical thinking skills. Thus, the designed numeracy questions must refer to critical thinking indicators integrated with the cognitive levels in the AKM model numeracy questions.

The next step is the concept analysis stage. In this stage, the determination of the domain developed in numeracy questions is done, aligning with the AKM 2021 framework. The content includes number and algebra, covering three contexts: personal, socio-cultural, and scientific, based on education for sustainable development. Based on the analysis results from previous activities, the purpose of developing numeracy questions based on education for sustainable development is to measure critical thinking skills and serve as a reference for numeracy questions.

The second main step is the design, starting with the development of a reference test, involving a literature review on ESD-based numeracy questions to measure critical thinking skills. This includes seeking information about global conditions, especially in Indonesia, related to sustainability aspects. The activity then proceeds with the development of a numeracy question matrix. The detailed components of AKM and ESD topics used in the questions are presented in Table 2.

Table 2. Distribution of question components

Content	AKM Components				Number of Question
	Cognitive Level	Context	ESD Topics	Question Format	
Number	Application	Scientific	Environment	Complex Multiple Choice	2
				Short Answer	1
	Reasoning	Social-Cultural	Economy	Complex Multiple Choice	3
				Short Answer	2
Algebra	Application	Scientific	Environment	Matching	1
				Complex Multiple Choice	1
	Reasoning	Personal	Environment	Multiple Choice	2
				Multiple Choice	1
Application	Scientific	Economy	Short Answer	1	
			Complex Multiple Choice	1	
Reasoning	Social-Cultural	Social	Complex Multiple Choice	1	

The next step is the selection of media. The chosen test medium is through Google Forms (g-form). This is done with the principle of sustainability, reducing the use of paper. However, the stimuli remain in hardfile form with the consideration that students can more easily work on the questions. The selected format is closely related to the previous step, which is the selection of media. In this research, the format referred to is the form of the question. The developed numeracy questions come in multiple-choice, complex multiple-choice, matching, and short answer formats. Questions in these forms are designed in such a way that students need the ability to analyze the given information, evaluate the answer options, and have a deep understanding of the content, as well as the ability to assess the accuracy or relevance of information. Numeracy questions also need to be accompanied by instructions to guide students before working on the questions.

In the initial design step, the process involves designing the numeracy test instrument, including the numeracy AKM matrix, answer alternatives, scoring guidelines, and numeracy questions adapted to the developed format. After the creation of the questions is completed, which in this case is in the form of prototype 1, the validation sheet is prepared before the test.

The third main step is development, starting with the validation process. The education for sustainable development-based numeracy questions were validated by two validators from the Mathematics Education Department at the University of Jember. The validation process of the research instrument is carried out by providing validation sheets along with matrices, education for sustainable development-

based numeracy question sheets, answer alternatives, and scoring guidelines. Aspects of expert validation for education for sustainable development-based numeracy questions consist of four aspects: material, construction, language, and ethics.

Based on the validation results attached, the average total score (V_a) for all aspects from both validators is 4.685, falling within the range of $4 \leq V_a < 5$. Therefore, the validity criteria for the developed numeracy questions are valid with revisions. After receiving and considering the suggestions from both validators, the revised version of the numeracy questions based on education for sustainable development is in the form of prototype 2.

The next step is the product testing phase, consisting of limited testing and field testing. The research instrument in the form of prototype 2 is tested in a limited manner with 7 randomly selected students from grade 12 at Senior High School in Bondowoso, Indonesia. The limited testing aims to assess the readability of the questions within a 60-minute timeframe. Meanwhile, field testing is conducted with students from grade 11 at Senior High School in Bondowoso. The purpose of the field testing is to measure the validity and reliability of the developed numeracy questions in assessing critical thinking abilities. Before answering the questions, respondents are provided with instructions and guidelines for responding to numeracy questions based on ESD.

The readability of the questions is analyzed based on a questionnaire on the readability of questions, considering feedback from limited testing subjects. The results of the limited testing respondents' questionnaire can be seen in Figure 1.

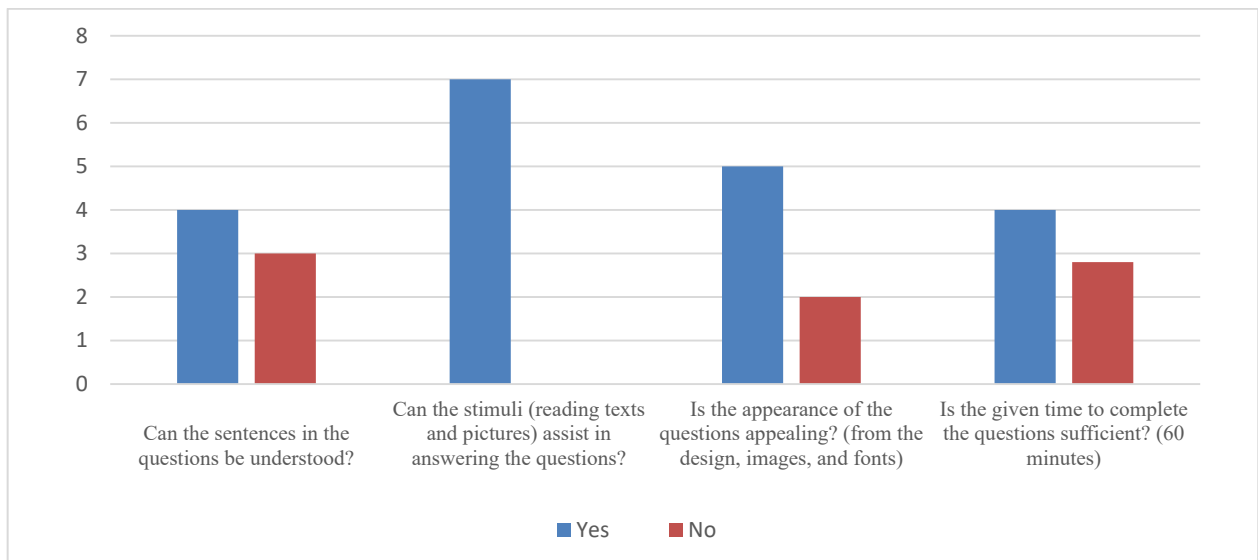


Figure 1. Results of respondent questionnaire as limited testing subjects

Based on the questionnaire results, there was a negative response from one respondent, indicating the need for a second limited test with a revised instrument, namely Prototype 2. The revision of the question instrument in the form of Prototype 2 was carried out, taking into account the feedback from

respondents regarding the developed numeracy questions.

The subsequent revision results were used for the second limited test with the same research subjects. The respondent questionnaire results from the second limited test are presented in Figure 2.

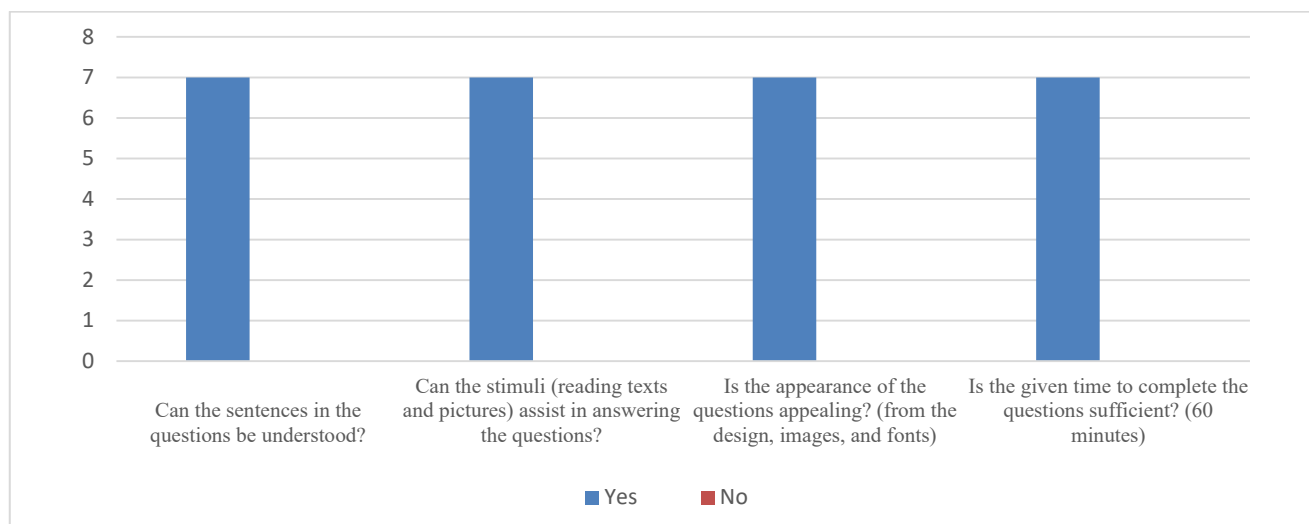


Figure 2. Results of respondent questionnaire as second limited testing subjects

Based on these questionnaire results, there were no negative responses from the respondents, indicating that the field trial could proceed. However, there were revisions or suggestions for question number 2 to provide more comprehensive information from the Ministry in Indonesia. Subsequently, the revised results of the numeracy question instrument based on Education for Sustainable Development in the form of Prototype 3 are presented in the QR Code in Figure 3.



Figure 3. Numeration problems based on ESD

Recapitulation of scores obtained by each respondent was conducted after the implementation of the field trial. The results of the scores in the field trial are raw scores that need further analysis. The raw scores of respondents will be processed using Ministep software. Each item is labeled S1, S2, and so on, up to S15, according to the order of the questions during the field trial. Rasch analysis with Ministep can depict the distribution of subject abilities and the distribution of item difficulty levels on the same scale. The item map can be seen in Figure 4.

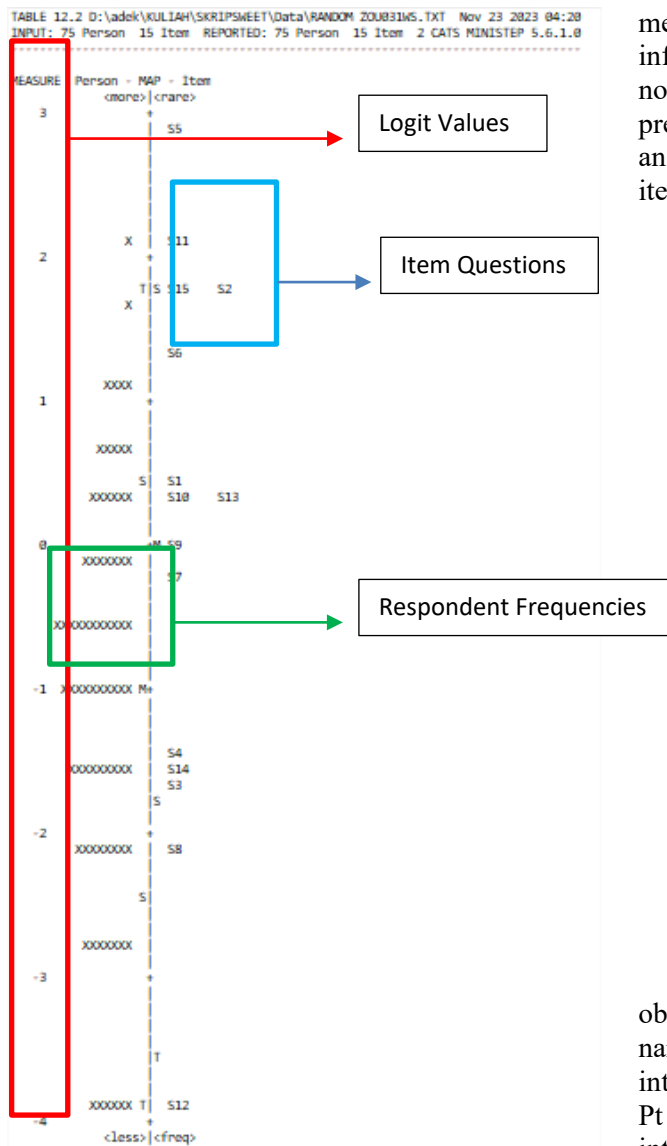


Figure 4. Item map

The distribution of subject abilities is on the left side, while the distribution of items is on the right side. Respondent frequencies indicate the spread of respondents within the logit value ranges, as well as with the items. For example, S5 is a question with the highest logit value, falling within the 2 to 3 intervals.

Generally, questions in the test are more difficult when compared to the subjects' abilities based on the item map. Theoretically, no subjects would have a chance of answering correctly on a question with code S5 because they have lower abilities than the difficulty level of that question.

3.1. Validity Analysis

The validity of an item can be examined based on three criteria: outfit mean square (MNSQ), outfit z-standard (ZSTD), and point measure correlation (Pt Measure Corr). An item is considered fit or valid if it meets at least one criterion. Item fit provides information on whether the developed item functions normally in measurement. An unfit item indicates the presence of misconceptions among subjects in answering that question. Information regarding the item fit of a particular item can be seen in Table 3.

Table 3. The results of the fit order items

No	MNSQ	ZSTD	Pt Mean Corr
S1	2.41	1.90	.21
S2	1.47	1.95	.43
S3	1.36	.66	.09
S4	1.36	.70	.25
S5	1.33	.96	.29
S6	1.28	1.06	.39
S7	.87	.06	.25
S8	.94	-.04	.50
S9	.80	-.87	.58
S10	.77	-.77	.55
S11	.73	-.75	.55
S12	.77	-1.04	.62
S13	.81	-.67	.64
S14	.31	-.93	.46
S15	.29	-1.08	.53

Based on the information in Table 3, it is obtained: Question item S6 meets one criterion, namely the outfit ZSTD, as it has a value within the interval -2 to 2, while the values of outfit MNSQ and Pt Measure Corr for S6 are outside the specified interval. Question items S7, S10, S12, and S15 meet two criteria, namely outfit MNSQ as they have values within the interval 0.5 to 1.5 and outfit ZSTD as they have values within the interval -2 to 2, while the Pt Measure Corr values for S7, S10, S12, and S15 are outside the specified interval. Question items S2 and S11 meet two criteria, namely outfit ZSTD as they have values within the interval -2 to 2 and Pt Measure Corr as they have values within the interval 0.4 to 0.85.

The outfit MNSQ values for S2 and S11 are outside the specified interval, while question items S1, S3, S4, S5, S8, S9, S13, and S14 meet all three validity criteria as they are within the specified intervals, i.e., outfit MNSQ values between 0.5 and 1.5, outfit ZSTD values between -2 and 2, and Pt Measure Corr values between 0.4 and 0.85.

These results indicate that all question items have met at least one validity criterion, and thus, all question items are considered fit or valid. The conclusion from this validity analysis is that there is no need to modify or eliminate any questions.

3.2. Reliability Analysis

The analysis results based on summary statistics from the Ministep software for field trial data involving 75 students as research subjects answering 15 numerical questions are presented in Table 4.

Table 4. The summary statistics results of the field trial

Measured Person	Separation	Reability
	1.57	.71
Item	4.85	.96
Alpha Cronbach	1.47	.71

Based on the information in Table 4, it is obtained that the Cronbach's alpha value, indicating the reliability measure, which is the interaction between respondents and test items, is 0.71, signifying good reliability. The person reliability value in the model is 0.71 with a separation of 1.57, meaning that the subjects are quite diverse as they have a wide range of abilities. The test items in the model have a separation of 4.85 and item reliability of 0.96. These values indicate that the quality of the questions is outstanding, and the test functions quite well, as it has a diverse range of difficulty levels, making the developed questions suitable for measuring students' abilities.

3.3. Difficulty Level Analysis

The results of the item difficulty level analysis are presented in Table 5 below.

Table 5. Measurement order of each test item

No	JMLE Measure	No	JMLE Measure
S5	2.92	S9	-.05
S11	2.11	S7	-.21
S2	1.74	S4	-1.49
S15	1.74	S14	-1.57
S6	1.31	S3	-1.65
S1	.46	S8	-2.06
S10	.37	S12	-3.88
S13	.28		

According to Table 5, item 5 has the highest logit value, which is +2.92, indicating that item number 5 is a question that few respondents can solve, specifically 3 out of 75 respondents. Meanwhile, item 12 has the lowest logit value, which is -3.88, signifying that item number 12 is a question that can be solved by many respondents, specifically 67 out of 75 respondents.

The difficulty level of the questions can be observed in the measure values. Following the guidelines for assessing items [21], questions 2, 5, 6, 11, and 15 fall into the category of very difficult questions as they have measures greater than +1. Questions 1, 10, and 13 are categorized as difficult questions because their measures are within the interval of 0 to +1. Questions 7 and 9 are categorized as easy questions as their measures fall within the interval of -1 to 0, while questions categorized as very easy are represented by questions 3, 4, 8, 12, and 14 due to having measure values below -1.

3.4. Student Ability Analysis

Information about the logit values of each respondent is presented in Table 6 below.

Table 6. Respondents' logit values

Respondent	JMLE Measure	Respondent	JMLE Measure	Respondent	JMLE Measure
A1004	2.13	A2001	-0.57	C1003	-1.54
C1006	1.61	A3005	-0.57	C1009	-1.54
A2006	1.15	A3006	-0.57	D1004	-1.54
A2007	1.15	A3008	-0.57	D1007	-1.54
C1004	1.15	A4002	-0.57	A1001	-2.11
C1010	1.15	A5009	-0.57	A4004	-2.11
A2002	0.71	A6004	-0.57	A5007	-2.11
A2003	0.71	A6007	-0.57	A5008	-2.11
A2005	0.71	C1008	-0.57	A6001	-2.11
B1006	0.71	D1003	-0.57	A6005	-2.11
C1002	0.71	A3002	-1.04	A6008	-2.11
A1003	0.29	A3004	-1.04	D1002	-2.11
A4005	0.29	A3007	-1.04	A6002	-2.83
B1002	0.29	A4001	-1.04	A6003	-2.83
B1004	0.29	A5002	-1.04	B1008	-2.83
B1007	0.29	B1003	-1.04	C2001	-2.83
C1007	0.29	C1005	-1.04	D1001	-2.83
A1006	-0.14	C2002	-1.04	D1006	-2.83
A3001	-0.14	C2003	-1.04	D1008	-2.83
A3003	-0.14	C2004	-1.04	A5001	-3.89
A4003	-0.14	A1005	-1.54	A5003	-3.89
A4006	-0.14	A2004	-1.54	B1001	-3.89
A5004	-0.14	A5005	-1.54	B1005	-3.89
A5006	-0.14	A6006	-1.54	D1005	-3.89
A1002	-0.57	C1001	-1.54	D1009	-3.89

The 'measure' column indicates the ability of students who were subjects in the study. Student A1004, a respondent ranked 4th from Group A1, has the highest logit value, which is +2.13.

This means that this student answered more questions correctly than the others. On the other hand, the lowest logit value, which is -3.89, is shared by 6 students with codes A5001, A5003, B1001, B1005, D1005, and D1009. This implies that these students answered more questions incorrectly than their peers.

The separation value from the Rasch model analysis, as indicated in Table 4, is 1.57, resulting in an H value as follows:

$$H=[4 \times \text{SEPARATION} + 1] / 3$$

$$H=[4 \times 1,57 + 1] / 3$$

$$H=2,42666\dots$$

The value is rounded up to 3. This means that Rasch models the measured abilities into three groups with different abilities: Level 1, Level 2, and Level 3. The logit value range is divided into three groups with the same logit interval. The frequency and percentage of each group in the field trial can be seen in Table 7.

Table 7. Frequency and percentage of student ability groups

Ability Group (P)	Logit Interval	Frequency	Percentage
Level 1	-3,89 ≤ P < -1,88	21	28%
Level 2	-1,88 ≤ P < 0,12	37	49%
Level 3	0,12 ≤ P ≤ 2,13	17	23%

Another output from the Rasch analysis in the Ministep software is the scalogram. The scalogram ranks student abilities from highest to lowest, while questions from very easy to very difficult are shown from left to right. In the annexed scalogram, it can be observed that the most difficult question can be answered by one student from level 3 ability group, i.e., student A1003, and two students from level 2 ability group, i.e., students A5009 and A3002. Students A5009 and A3002 can answer the most difficult question but cannot correctly answer other difficult and very difficult questions. This might suggest that these students guessed the answer, considering that question number 5 is a reasoning question that requires in-depth analysis and determining the correct informed argumentation.

4. Discussion

The development process of numeracy questions has gone through the stages of definition, design, and development. The developed numeracy questions are aligned with the characteristics of AKM questions and critical thinking indicators as defined in the initial stage. There are 15 questions that were then tested on 75 respondents to determine the quality criteria of the developed questions.

The quality of the developed questions is assessed based on the results of validity and reliability analyses. Validity measures the extent to which test items truly measure what is intended to be measured, namely critical thinking. Reliability measures how consistently test items can produce the same scores when measured at different times or in different ways.

The field trial results were analyzed using the Ministep software with Rasch modeling. Educational assessment and evaluation become more objective with the Rasch model, including through the item map feature that can show whether the developed test accommodates various levels of competence of the measured respondents [22]. The Rasch model will maintain the level of difficulty of questions invariant, regardless of the characteristics of the sample used in the initial validation.

Based on the item map results, the distribution of subject abilities and the difficulty level distribution show that the questions tend to be more difficult compared to the abilities of the subjects. The validity of each item is assessed based on three criteria: outfit mean square (MNSQ), outfit z-standard (ZSTD), and point measure correlation (Pt Measure Corr). These criteria look at the fit of the items predicted by Rasch and whether response string aligns with the Guttman-style model. The conclusion from the validity analysis is that all items meet at least one validity criterion, indicating that all questions can be considered valid and do not require modification or elimination. A valid test instrument means that all questions can be used to measure mathematical critical thinking abilities [20].

The Cronbach's alpha value is 0.71, indicating that the relationship between student responses to various test items has a good level of consistency. This reliability quality indicates that the interaction between students and the given questions can be relied upon to measure students' abilities overall. The person reliability value is 0.71 with a separation of 1.57, which can be interpreted as the subjects who took the test showed sufficient variation. This respondent reliability occurs because the number of subjects used is only 75 participants. The test item value has a separation of 4.62 and an item reliability of 0.96. With an item separation of 4.85 and item reliability of 0.96, it can be concluded that the quality of the test items is outstanding. This indicates that the test has successfully covered a diverse range of difficulty levels and ensures that the questions posed are suitable for measuring and assessing various student abilities [20].

Based on the analysis of the sequence of measurements for each item, there are four difficulty levels of questions.

Specifically, there are five questions classified as very difficult, three questions as difficult, two questions as easy, and five questions as very easy. The most challenging questions in this study assess reasoning abilities through complex multiple-choice questions where students cannot select all correct answer choices with appropriate reasoning. Students' ability to understand, transform, and process problem-solving will affect the conclusions made by students [23]. The success of students in solving numeracy questions also depends on the topic or stimulus provided. The low interest of students in reading can lead to a lack of motivation for students to seek information contained in the topic [24]. In addition, questions categorized as very difficult and difficult tend to involve complex calculations, so in the future, students are expected to improve their problem-solving skills and deepen their mastery of the material, partly through practice.

In the field trial of this study, the results indicate that 28% of students are in the level 1 ability group, 49% of students are in the level 2 ability group, and 23% of students are in the level 3 ability group. These results indicate that there are more students in the lower to middle-level ability group compared to those in the upper to middle-level ability group. This distribution may not meet the desired expectations. Teachers should focus on more inclusive strategies to assist students of various ability levels. Learning efforts can be focused on more differential methods, with more attention to students who need further assistance. Differentiated learning in a flexible curriculum is crucial to responding to the diverse learning needs of students to create relevant learning experiences [25]. Moreover, learning strategies that stimulate and encourage students to develop their critical thinking skills optimally need to be improved, emphasizing the need for adaptive learning with ESD-based content. This is intended to promote interdisciplinary and holistic approaches and foster critical and creative thinking in the education process [7].

5. Conclusion

The development of numeracy questions based on education for sustainable development to measure critical thinking skills has gone through the stages of Thiagarajan's model, including define, design, and develop. In the definition stage, the researcher highlighted the need for improving critical thinking skills in Indonesia through numeracy questions based on ESD model AKM. In the design stage, the formulation of the numeracy question prototype was done using global information and literature to build the question framework.

In the development stage, field trials were conducted using the Ministep software for Rasch analysis. The results of the development of numeracy questions meet valid and reliable parameters with the details as follows: (1) All questions are declared valid as they have met at least one validity criterion, thus not requiring any changes or deletions; (2) an item reliability of 0.96 indicates that the quality of the test item is exceptional and has good ability in measuring students' critical thinking skills; (3) Numeracy questions can measure critical thinking skills, with 28% of students included in level 1 ability group, 49% in level 2, and 17% in level 3. These results indicate that the ESD-based numeracy problems developed can be used as reference material in similar research or studies related to the analysis of students' critical thinking abilities in solving ESD-based numeracy problems.

References:

- [1]. Wagner, T. (2010). *Overcoming The Global Achievement Gap*. Cambridge, Mass: Harvard University.
- [2]. Kennedy, T.J., Sundberg, C.W. (2020). 21st Century Skills. In Akpan, B., Kennedy, T.J. (eds) *Science Education in Theory and Practice. Springer Texts in Education*. Springer, Cham. Doi: 10.1007/978-3-030-43620-9_32
- [3]. Ennis, R.H. (2015). *The Nature of Critical Thinking: Outlines of General Critical Thinking Dispositions and Abilities*. Education. Retrieved from: https://education.illinois.edu/docs/default-source/faculty-documents/robert-ennis/thenatureofcriticalthinking_51711_000.pdf [accessed: 15 April 2024].
- [4]. Alsaleh, N. J. (2020). Teaching Critical Thinking Skills: Literature Review. *Turkish Online Journal of Educational Technology-TOJET*, 19(1), 21-39.
- [5]. OECD. (2023). *PISA 2022 Results (Volume II): Learning During – and from – Disruption*. Paris: OECD Publishing. Doi: 10.1787/a97db61c-en.
- [6]. Monrat, N., Phaksunchai, M., & Chonchaiya, R. (2022). Developing Students' Mathematical Critical Thinking Skills Using Open-Ended Questions and Activities Based on Student Learning Preferences. *Education Research International*, 2022, 1-11. Doi: 10.1155/2022/3300363
- [7]. Vásquez, C., Alsina, Á., Seckel, M. J., & García-Alonso, I. (2023). Integrating sustainability in mathematics education and statistics education: A systematic review. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(11), 1-16. Doi: 10.29333/ejmste/13809
- [8]. York, T.T., Gibson, C., & Rankin, S. (2019). Defining and Measuring Academic Success. *Practical Assessment, Research, and Evaluation*, 20(5), 1-20. Doi: 10.7275/hz5x-tx03
- [9]. Butterworth, J. & Thwaites, G. (2013). *Thinking Skills Critical Thinking and Problem Solving Second Edition*. UK: Cambridge University Press.

- [10]. Gantiyani, H., Hobri, Cahya, P.A., Atika S.D., Hariati, A. (2022). Student's critical thinking ability in solving AKM numeration problems with three different cognitive levels. *AIP Conf. Proc.*, 2633(1): 030019. Doi: 10.1063/5.0109966
- [11]. Forgasz, H. J., & Hall, J. (2019). Learning about Numeracy: The Impact of a Compulsory Unit on Pre-service Teachers' Understandings and Belief. *Australian Journal of Teacher Education*, 44(2), 15-33. Doi: 10.14221/ajte.2018v44n2.2
- [12]. Hoogland K. (2023). The changing nature of basic skills in numeracy. *Front. Educ.* 8, 1293754. Doi: 10.3389/feduc.2023.1293754
- [13]. Jain, P., & Rogers, M. (2019). Numeracy as Critical Thinking. *Adults Learning Mathematics: An International Journal*, 14(1), 23-33.
- [14]. Taimur, S., & Sattar, H. (2020). Education for sustainable development and critical thinking competency. *Quality education. Encyclopedia of the UN Sustainable Development Goals*, 238-248. Doi: 10.1007/978-3-319-95870-5_64
- [15]. Machromah, I. U., Ishartono, N., Mirandhani, A., Samsudin, M., & Basry, W. (2021). PISA Problems Solving of Students with a Visual Learning Styles. *Journal of Physics: Conference Series*, 1720(1), 012010. Doi: 10.1088/1742-6596/1720/1/012010
- [16]. Sumarno, W. K., Shodikin, A., Solikha, N. I. A., Pratama, N. K., & Valensiana, B. F. (2022). Integrative Teaching Material with Project-based Learning Approach to Improve Elementary School Students' Bilingual Literacy and Numeracy Skills. *International Journal of Elementary Education*, 6(4). Doi: 10.23887/ijee.v6i4.52392
- [17]. Thiagarajan, S., Semmel, D.S., & Semmel, M.I. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Indiana: Indiana University Bloomington.
- [18]. Dinnesen, M. S., Oiszewski, A., Breit-Smith, A., & Guo, Y. (2020). Collaborating With an Expert Panel to Establish the Content Validity of an Intervention for Preschoolers with Language Impairment. *Communication Disorders Quarterly*, 41(2), 86-99. Doi: 10.1177/1525740118795158.
- [19]. Wahyuningsih, S. (2021). Using the Rasch's Partial Credit Model to Analyze the Quality of an Essay Math Test. In *1st International Conference on Mathematics and Mathematics Education (ICMMEd 2020)*, 257-265. Atlantis Press. Doi: 10.2991/assehr.k.210508.073
- [20]. Harvani, I. D., Kurniati, D., Kim, D. J., & Osman, S. (2023). Quality of Algebraic Numeration Problems to Measure Higher Order Thinking Skills Using Partial Credit Model. *The New Educational Review*, 72, 218-229.
- [21]. Takács, R., Kárász, J.T., Takács, S. et al. (2021). Applying the Rasch model to analyze the effectiveness of education reform in order to decrease computer science students' dropout. *Humanities and Social Sciences Communication*, 8(44), 1-8. Doi: 10.1057/s41599-021-00725-w.
- [22]. Wright B.D., Stone M.H. (1979). *Best test design - Rasch Measurement*. Mesa. Chicago: Mesa press.
- [23]. Zaini, A. H. & Retnawati, H. (2019). What Difficulties that Students Working in Mathematical Reasoning Questions?. In *Journal of Physics: Conference Series*, 1397, 1-9. Doi: 10.1088/1742-6596/1397/1/012079
- [24]. Barber, A. T., & Klauda, S. L. (2020). How Reading Motivation and Engagement Enable Reading Achievement: Policy Implications. *Policy Insight from the Behavioral and Barain Sciences*, 7(1), 27-34. Doi: 10.1177/2372732219893385.
- [25]. Eikeland, I., & Ohna, S.E. (2022). Differentiation in education: a configurative review. *Nordic Journal of Studies in Educational Policy*, 8(3), 157-170. Doi: 10.1080/20020317.2022.2039351