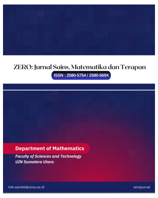

Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754
P-ISSN : 2580-569X
Vol. 8, No. 2, 2024
DOI: 10.30829/zero.v8i2.21807
pp. 90-98

# A Mathematics Learning Outcome Test Characterized by Higher-Order Thinking Skills for Grade V Elementary School Students

**I Komang Sesara Ariyana[1], Sariyasa[2]**
[1]Department of Dharma Acarya, Sekolah Tinggi Agama Hindu Negeri Mpu Kuturan Singaraja, Bali, Indonesia
[2]Mathematics Education, Universitas Pendidikan Ganesha, Bali, Indonesia

**Article Info**

***Article history:***

Received, 02 10 2024
Revised, 05 11 2024
Accepted, 10 12 2024

***Keywords:***

Higher-order thinking;
Math problems solving;
Multiple choice;
Test.

**ABSTRACT**

This research was aimed at developing the results of learning mathematics characteristics of higher-order thinking skills (HOTS) in the form of multiple choice for grade V elementary school students. This research applied D & D model by Richey & Kleine (2007). HOTS in this study were able to transfer, critical thinking, and solve problems. These aspects are included in Analyzing, Evaluating, and Developing category. Math questions which HOTS were Multi step problems and non-routine questions. The sample of the test experiment was 112 students drawn from four elementary schools in Singaraja Town. The results showed that of 21 items that had been made, there were 20 valid items. The difficulty analysis showed five unfit grains, 14 items were categorized as difficult and one item was categorized as moderate. By the discrimination index, there were two items categorized as high, nine items were categorized as moderate, and four items were categorized as low. The effectiveness of the distractors was fulfilled in all questions. These fifteen items have a degree of reliability of 0.609.



***Corresponding Author:***

I Komang Sesara Ariyana,
Department of Dharma Acarya,
Sekolah Tinggi Agama Hindu Negeri Mpu Kuturan Singaraja, Bali, Indonesia
Email: sesara.ariyana@stahnmpukuturan.ac.id

## 1. INTRODUCTION

Globalization, new technologies, migration, international competition, changing markets, and cross-border environmental and policy challenges push students to acquire all the skills and knowledge they have to survive and succeed in the 21st century [37]. In the context of school education, students need to learn to have essential skills to succeed in today's world, such as critical thinking, problem solving, communication and collaboration [28].

Mathematics is a core subject in the 21st century [28]. Mathematics is a common language that can help students solve complex problems and is a lens of understanding to make important connections to other fields, professions and disciplines. The ability to solve problems is the main goal of learning mathematics. Everyone needs to learn mathematical problem solving to be able to live in the 21st century productively (Holmes, 1995, in Wardhani, et al., 2010) [47]. NCTM (2000) also states that the goal of school mathematics should be for all students to become increasingly able and willing to engage in problem solving [23]. However, this goal of mathematics education does not seem to have been achieved optimally.The achievement of Indonesian students is considered not encouraging based on their participation in the international studies Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA) since 1999. This is due to, among other things, the large number of test materials asked in TIMSS and PISA not being included in the Indonesian curriculum (Regulation of the Minister of Education and Culture of the Republic of Indonesia No. 67 of 2013 concerning the Elementary School Curriculum) [30]. In addition, according to Lutfianto, et al. (2013), one of the factors that causes low-achieving classes is the habit of students in solving contextual problems at school [21]. Budiarti et al. (2023) also suspect that this is due to the lack of ability of Indonesian students in facing and solving mathematics problems [8]. Common problems in schools are different from problems on the PISA test because they use real-life situations.

Developing and improving students' higher-order thinking skills (HOTS) is one of the main goals of learning in the 21st century [15]. Higher-order thinking skills are important for learners to master, because they can motivate learners to look at every problem critically, creatively, logically, and objectively [25]. The Common Core State Standards Initiative (CCSSI, in Partnership for 21st Century Skills, 2008) also explicitly calls for, and integrates, higher-order thinking skills (HOTS) as a means to achieve career and college readiness for all students [29]. HOTS is essential for survival and success in the 21st century [37]. Nizam (2016, in Widana, 2017) stated that assessment in Indonesia is directed at the higher order thinking skills (HOTS) assessment model [48]. This policy refers to the need for life skills in the 21st century. But Ariyana (2020) showed that 72% of teachers had never created HOTS-characterized math problems, 72% did not have examples of HOTS problems, and 45% of teachers did not like practicing HOTS-characterized math problems [5]. In fact, practicing HOTS questions is important for better thinking skills that can help solve many problems in everyday life.

There is a tendency that HOTS questions are difficult questions. It is important to realize that the level of difficulty (easy – hard) and the level of thinking (LOTS – HOTS) are two different qualities [7]. The misconception that remembering is "easy" and high-level thinking is "hard" can lead to poor results.

Brookhart (2010) stated that multiple-choice questions, especially those that use an introduction, can also assess HOTS [7]. What needs to be considered are the principles of higher-order thinking assessment when writing the items or tasks to be given, namely (1) using introductory materials, (2) using new materials, and (3) separating cognitive levels from levels of difficulty [7]. According to Scully (2017), multiple-choice questions that measure complex cognitive processes are rarely created [38]. In other words, HOTS measurement is not determined by the form of the test, but rather how the measurement (assessment) can encourage higher-order thinking from test takers (students) in accordance with these principles.

For multiple-choice test scores to be generated that mean that students are using higher-order thinking, the questions must be designed in such a way that higher-order thinking is actually required to answer [7]. Widana (2017) also argues that in the form of multiple-choice questions, the answer choices represent the ability of students to solve problems [49]. According to Rubin & Rajakaruna (2015), the potential to build multiple-choice questions to evaluate higher-order thinking has now evolved [34].

The purpose of this study was to develop a multiple-choice mathematics learning outcome test. The grand theory used for the development of this multiple-choice test uses Classical Test Theory (CTT). Classical test theory (CTT) is a traditional quantitative approach used to test the reliability and validity of an instrument based on the items it contains [27]. This study is based on the results of Ariyana's (2020) study which showed that grade V teachers in Singaraja City, Buleleng Regency, Bali Province needed a HOTS-characterized elementary school mathematics learning outcome test [5]. In addition, the results of Sumarni's (2022) study showed that the ability of teachers to compile HOTS questions at SDN 02 Legokgunung was good, especially in the construction aspect, even very good in the language aspect, but still weak in the material aspect, especially in questions measuring the cognitive level of reasoning (analyzing, evaluating, creating) [43].

Higher Order Thinking Skills (HOTS) or high-level thinking skills in this study refer to the theory put forward by Brookhart (2010) that HOTS is being able to transfer, being able to think critically, and being able to solve problems [7]. In terms of transferring, Brookhart (2010) refers to the Revised Bloom's Taxonomy by Anderson and Krathwohl (2001) that HOTS is the top three levels of cognitive processes in the taxonomy are HOTS, namely Analyzing, Evaluating, and Creating [2]. Associated with mathematics, the three levels of cognitive processes are associated with mathematical problem-solving questions. These mathematical problem-solving questions were created by referring to Van de Walle et al. (2013) and Sonnabend (2010), where the study provides direction that HOTS mathematics questions are procedural questions with connections and mathematical work questions [41][46]. Procedural questions with connections in this case are multi-step questions, while mathematical work questions are non-routine questions [41].

## 2. RESEARCH METHODE

This research was a design and development (D&D) by Richey & Kleine (2007). Design and development research in this context is a systematic study of the design, development, and evaluation process that aims to build an empirical basis for creating learning and nonlearning products and tools, as well as new or improved models that serve as references in their development [33]. The research method used was Design and Development (D&D) with three main elements: design, development and evaluation

[20]. The product produced in this study is a Mathematics Learning Outcome Test with Higher-Order Thinking Skills (HOTS) Characteristics for Grade V Elementary School Students.

At the Design Stage, namely determining the provisions of a good test in making the initial design of the test. The steps in this design stage are (1) specifying learning objectives; (2) creating the test blueprint; (3) format selection and determination; and (5) establishing good test criteria. Furthermore, at the Development stage, it is carried out in terms of making questions based on the conceptual framework of development and validating the content by experts. The characteristics of the questions from the developed test are (1) using introductory material, (2) using new material, and (3) using a higher cognitive level (analyzing, evaluating, and creating based on Revised Bloom's Taxonomy) that differs from the level of difficulty.

Testing the validity of the content was carried out using the Lawshe technique (1975, in Shultz, et al., 2014) with three experts in the field of mathematics education [340. The characteristics of the CVR formula according to Lawshe (1975, in Nengsih et al, 2019) are: 1) when all panellists answered 'important', the CVR value is 1. When more than half of the panellists answered 'important', but not all of them, the CVR value ranges from 0 to 0.99; 2) when less than half of the panellists answered 'important', the CVR value will be negative; and 3) when half of the panellists answered "important" and half answered 'not important', the CVR value is zero [24].

At the Evaluation stage, limited trials and empirical validation were conducted. A total of 112 fifth grade students were taken from 4 Elementary Schools in Buleleng District who were taken randomly using the cluster random sampling technique. Empirical validation of the test was conducted by testing the validity of the test items, analyzing the level of difficulty of the test items, analyzing the discriminating power of the test items, analyzing the effectiveness of the distractor items, and testing the reliability of the test.

Testing the validity of the test items was conducted using the point biserial correlation technique ($\gamma_{pbi}$) with a critical r = 0.30. Analysis of the level of difficulty of the test items was conducted by referring to the Davis index. A test item can be used if the item has an average difficulty index between 0.25 and 0.75 [10]. Analysis of the discriminating power of the test items was conducted by referring to the Flanagan table which uses a biserial coefficient that moves from 0.20 to 0.80 [10]. Analysis of the effectiveness of the test distractors was conducted to check whether the distractors met the criteria that a good distractor is chosen by more than 5% of test participants. Finally, reliability testing was conducted using the KR-20 formula ($r_{11}$) with a degree of reliability of more than 0.60.

## 3. RESULT AND ANALYSIS

This research is a continuation of Ariyana's study (2020) research on the needs analysis of mathematics tests with HOTS-characterized for Grade V elementary school students in Buleleng District [5]. The results of the study at need analysis are in accordance with Ariyana's study (2020) which showed that 72% of teachers had never created HOTS-characterized math problems, 72% did not have examples of HOTS problems, and 45% of teachers did not like practicing HOTS-characterized math problems [5]. As the grade V mathematics teachers in these primary schools lack competency development experiences, it suggested that there was a need for HOTS-characterized mathematics tests for grade V students in Buleleng District. This finding also supports the results of previous research conducted by Apino & Retnawati (2017) that teachers' understanding of HOTS is also lacking, and there are even high school math teachers who are not familiar with the term HOTS [3]. This finding was also confirmed by Putra (2016) that the evaluation system in Indonesia still uses low-level questions, and students are accustomed to obtaining and using formal mathematical knowledge in class [32]. Students will not experience high-level thinking if teachers never construct HOTS questions [5].

At the Design stage, test design is carried out by referring to qualitative and quantitative criteria. The specifications of the mathematics learning outcome test characterized by higher-order thinking skills are in accordance with the conceptual framework of development. Each question is made with reference to the Revised Bloom's Taxonomy, namely at the cognitive process level of Analyzing (C4), Evaluating (C5), or Creating (C6). The abilities that appear in these categories are being able to transfer, think critically, and solve problems. In addition, the questions that have been made are also divided into multi-step questions and non-routine questions as HOTS-characterized mathematics questions. Referring to the Revised Bloom's Taxonomy, the questions made are in the scope of Conceptual Knowledge (K2) and Procedural Knowledge (K3). There are 10 questions at the Analyze level. There are 6 questions at the Evaluate level. There are 5 questions at the Create level. There are 13 questions with the multi-step problem category, and 8 questions with the non-routine problem category. The Develop stage is an activity to create a HOTS-characterized mathematics learning outcome test and validation activities. Validation at the Develop stage is content validity which includes expert validity.

The learning outcome test was created by referring to the fifth-grade elementary school material, namely the Fraction Arithmetic Operations and Comparison and Scale material. The number of questions that had been created at the time of the initial design was 21 items with details of the Fraction Arithmetic Operations material as many as 10 questions, while the Comparison and Scale material as many as 11 questions. NCTM (2000) directs that good problems will integrate several topics and will involve significant mathematics [23]. Therefore, HOTS characteristic questions require complex cognitive processes and knowledge in solving problems.

Each item represents one HOTS category, namely questions with cognitive process levels in the Analyzing (C4), Evaluating (C5), and Creating (C6) categories. HOTS-characterized mathematics questions that have been created in this study are similar to the categorization of the three complex cognitive processes by Tajudin (2015), who distinguishes the three HOTS categories in mathematics learning, where when students identify the necessary elements and determine the relationship between these elements is a process that requires the ability to Analyze; when students translate text and diagrams into symbolic representations (equations), they involve the ability to Create; and when students are asked to engage in activities that check the reasonableness (of a statement or calculation result) requires the ability to Evaluate [44].

Testing the validity of the content of the interest aspect using the CVR technique (Content Validity Ratio) by Lawshe (1975, in Shultz, et al., 2014) [40]. Content validity testing was conducted by three experts in the field of mathematics education. The three experts in this study labeled each question that had been created as "important" so that the CVR value for each question was 1. Gilbert & Prion (2016) emphasized that if all panelists agreed that an item was "important", the CVR was 1.00 [18]. So, all questions created were included in the HOTS-characterized learning outcome test to be tested immediately.

In the final stage, namely Evaluation, limited trials and empirical validation were conducted. The questions that have been developed may be used by others to train students to improve HOTS in daily learning. As stated by Arafah, et al. (2021), that the HOTS instrument assessment as an effective learning assessment to train students' HOTS and is effective in measuring students' thinking skills based on each student's HOTS level [4].

By using the formula manually and calculating it using Ms EXCEL, the validation results can be seen in Table 1 below. The data displayed in Table 1 are the results of content validity, item validity testing using the point biserial formula, analysis of the difficulty level of the questions using the average of respondents who answered correctly and the Davis index [13], and analysis of the question discrimination index using the formula d = (U - L)/N [27].

**Tabel 1.** Summary of Validation Results

| No. Question | Dimension | Problem Category | Validity of Content | | Item Validity ($r_{critical}$= 0.30) | | Difficulty Level | | Discrimmination Power | | Final decision Status |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *CVR* | Note | $r_{count}$ ($\gamma_{pbi}$) | Status | *D* | Statuts | $r_{bis}$ | Statuts | |
| 1 | C4, K3 | Multi-step | 1 | Good | 0.328 | Valid | - | *Not feasible* | 0,3 | Satisfactory | Thrown |
| 2 | C4, K3 | Multi-step | 1 | Good | 0.336 | Valid | 65,018 | High | 0,5 | Good | Used |
| 3 | C6, K3 | Multi-step | 1 | Good | 0.401 | Valid | 70,347 | High | 0,63 | Good | Used |
| 4 | C4, K3 | Multi-step | 1 | Good | 0.398 | Valid | 72,474 | High | 0,4 | Good | Used |
| 5 | C6, K3 | Multi-step | 1 | Good | 0.390 | Valid | - | *Not feasible* | 0,27 | Satisfactory | Thrown |
| 6 | C6, K3 | Multi-step | 1 | Good | 0.369 | Valid | 69,357 | High | 0,33 | Satisfactory | Used |
| 7 | C5, K2 | Non-routine | 1 | Good | 0.323 | Valid | 70,347 | High | 0,4 | Good | Used |
| 8 | C5, K2 | Non-routine | 1 | Good | 0.385 | Valid | - | *Not feasible* | 0,37 | Satisfactory | Thrown |
| 9 | C4, K3 | Multi-step | 1 | Good | 0.410 | Valid | - | *Not feasible* | 0,5 | Good | Thrown |
| 10 | C6, K3 | Multi-step | 1 | Good | 0.361 | Valid | 77,614 | High | 0,37 | Satisfactory | Used |
| 11 | C4, K3 | Multi-step | 1 | Good | 0.223 | *Invalid* | - | - | - | - | Thrown |
| 12 | C5, K2 | Non-routine | 1 | Good | 0.314 | Valid | 73,633 | High | 0,5 | Good | Used |
| 13 | C4, K3 | Non-routine | 1 | Good | 0.304 | Valid | 73,633 | High | 0,4 | Good | Used |
| 14 | C4, K3 | Multi-step | 1 | Good | 0.321 | Valid | 74,854 | High | 0,3 | Satisfactory | Used |

| No. Question | Dimension | Problem Category | Validity of Content | | Item Validity ($r_{critical}$= 0.30) | | Difficulty Level | | Discrimmination Power | | Final decision Status |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $CVR$ | Note | $r_{count}$ ($\gamma_{pbi}$) | Status | $D$ | Statuts | $r_{bis}$ | Statuts | |
| 15 | C5, K3 | Multi-step | 1 | Good | 0.412 | Valid | 74,854 | High | 0,47 | Good | Used |
| 16 | C5, K3 | Non-routine | 1 | Good | 0.335 | Valid | 72,474 | High | 0,3 | Satisfactory | Used |
| 17 | C4, K3 | Non-routine | 1 | Good | 0.352 | Valid | 73,633 | High | 0,37 | Satisfactory | Used |
| 18 | C6, K3 | Non-routine | 1 | Good | 0.353 | Valid | 68,493 | High | 0,4 | Good | Used |
| 19 | C4, K3 | Multi-step | 1 | Good | 0.369 | Valid | 74,854 | High | 0,37 | Satisfactory | Used |
| 20 | C4, K3 | Multi-step | 1 | Good | 0.467 | Valid | - | *Not feasible* | 0,53 | Good | Thrown |
| 21 | C5, K3 | Non-routine | 1 | Good | 0.582 | Valid | 53,181 | Currently | 0,6 | Good | Used |

From the results of the validity test of the test containing 21 questions in it, it was obtained that 1 question was invalid, namely number 11. Question number 11 is a question taken from KD 3.2, namely explaining and multiplying and dividing fractions and decimals. This question is included in the category of Analyzing Procedural Knowledge (C4, K3).

Based on the results of calculating the level of difficulty of each question item that has been declared valid, it was found that there were 5 questions that were not used because the D value was not obtained. This happened because the Pb value was 0.000 or negative. Of the questions that could be used, 14 questions had a high level of difficulty and 1 question had a medium level of difficulty. Questions with high difficulty may be caused by students not understanding the meaning of the question, the use of words and terms that students do not understand, or even students are not used to complex story-based questions. The language aspect is indeed a problem for students in answering story questions. This is in line with the opinion of Shete & Kausar (2015) based on their findings which state that this may be caused by poor understanding of difficult topics, ambiguity in various question words or inappropriate key or personal variations in forming multiple choices [39]. In fact, according to Zahra & At-Taqiyyah (2024), high-level multiple-choice tests have the potential to measure in-depth understanding, encourage critical thinking, develop analytical skills, encourage problem solving, and prepare for standardized testing [51].

Initially, the researcher's aim was to develop multiple-choice questions whose answer choices were statements. This is supported by Scully (2017) who in his article provides suggestions for several strategies that can be done in creating multiple-choice questions that measure HOTS [38]. One of them is using high-quality distractors. The point is that if possible, all the choices given are theoretically reasonable, with the key being the "best" answer, as opposed to the only correct choice. The stem must also be given the right words to reflect this.

In addition, another possibility is that students are not yet accustomed to complex story-based problems supported by Wijaya, et al. (2015) who found that the number of context-based tasks in textbooks in Indonesia is low [50]. When solving context-based tasks, students have difficulty in (1) understanding what the problem is, (2) distinguishing between relevant and irrelevant information, and (3) identifying the mathematical procedures needed to solve a problem [50].

Students' mathematical problem-solving ability is still weak. The weak mathematical problem-solving ability of students is caused by the problem factors and learning factors. From the problem factors, Ruhyana (2016) found that students had difficulty in understanding problems in story problems and interpreting them into mathematical sentences [35]. Students are also not yet familiar with problem-solving questions [16][42].

From the learning factors, it was found that teachers found it difficult to teach mathematical problem-solving skills [31], teacher-centered learning [17], and teachers were less able to design learning [14]. Usmaedi (2017) showed that learning in elementary schools so far tends to emphasize memorization aspects alone, without being followed by deep understanding and comprehension [45]. Santi et al. (2021) summarized several problems faced in learning mathematical problem solving in elementary schools that can also originate from teachers, where (1) teachers do not emphasize problem solving enough; (2) teachers experience misconceptions related to problem solving questions; (3) teachers experience difficulties in teaching problem solving; (4) teachers still apply a teacher-centered learning model; and (5) teachers' wrong perceptions about problem solving and learning have implications for their learning [36].

In addition, for the difficulty in working on fractional arithmetic operations, Van de Walle, et al. (2013) emphasized that fractional notation contributes to students' difficulties with fractions. Aksu (1997, in Wijaya,

2017) also stated that in relation to fractional operations, students perform equally well with addition, subtraction, multiplication, and division when the task is presented as a calculation [49]. However, when the task is given in the form of a story problem, addition is the easiest and multiplication is the most difficult.

Most students fail to solve fraction problems involving the use of HOTS [1]. The failure may be caused by the teaching by teachers who do not emphasize the understanding and mathematical skills needed by students and ultimately affect them because they fail to produce the desired results. In short, students need to equip themselves with various high knowledge and skills in problem solving, which involves high-level thinking [1].

There are two reasons why high difficulty items are still used. First, the items are not very difficult. Second, very difficult items do not need to be dropped or discarded. The first reason is supported by Menon & Kannambra (2017) who emphasized that difficult questions can be maintained and used to select students with the highest rankings [22]. Very difficult items can still be used if teachers want to select students with high achievement. The second reason is supported by Shete & Kausar (2015) who state that very difficult and very easy items do not need to be dropped, but rather need to be reconstructed and reassessed [39].

The main determinant of the quality of multiple-choice items is their discriminatory power (Ebel, 1975, in DiBattista & Kurzawa, 2011), which reflects the extent to which more knowledgeable students are more likely than less knowledgeable students to select the key option [11]. For discriminatory power, the higher the discriminatory power index, the better the item can differentiate between high-scoring and low-scoring students [39]. Based on the results of the discriminating power test, it can be seen that of the 15 questions that were retained, there were 4 questions with low discriminating power, 9 questions with medium discriminating power, and 2 questions with high discriminating power. The questions with low discriminating power were still used becauseit alerts us to the possibility of technical defects in test items but should not cause us to discard items that are considered valuable. A well-constructed achievement test will by necessity contain items with low discriminating power and to discard them would result in a test that is less, not more, valid [19].

Before testing the reliability of the test, for multiple-choice questions, it is necessary to analyze the effectiveness of distractors. In this study, each multiple-choice question provided four answer choices with one answer key and three distractors. The number of participants in the trial test was 112 students. This means that 5% of 112 is 5.6 which is rounded up to 6. Based on the results of the analysis, all questions had distractors that functioned well, where all distractors were chosen by more than 6 students.

DiBattista & Kurzawa (2011) stated that the discriminatory power of multiple-choice questions is highly dependent on the quality of the distractors [11]. In addition, for HOTS-measuring questions as stated by Scully (2017) that distractors that are superficially similar to the answer key, on the other hand, require a high level of discriminatory power assessment [38]. Scully (2017) suggested that if possible, all choices given are theoretically plausible, with the key being the “best” answer, as opposed to the only correct answer choice [38].

Next, the reliability test of the fifteen questions. From the results of the test reliability test of 15 questions, it was obtained that the test had a reliability degree of 0.609. According to Guilford (1951, in Candiasa, 2010), this means that this test has high reliability or can be said to be reliable [9]. However, according to El-Uri & Malas (2013), a KR20 figure of 0.8 is considered the minimum acceptable score [12]. A figure below 0.8 can indicate various events, namely that the test is very difficult, or perhaps it tests unknown or unexpected topics. KR20 is influenced by the difficulty, distribution of scores, and length of the test. As stated by Gronlund (1981) that the factors that influence reliability are the length of the test, distribution of scores, difficulty of the test, reliability estimation method, and objectivity [19].

Gronlund (1981), stated that in general, the longer the test, the higher the reliability [19]. Nunnally (1978) argued that usually a minimum of 30 items have high reliability [26]. While in this study only produced 15 items that were considered good. This means that it is necessary to add items until the reliability is as high as desired [26]. The greater the spread of scores, the higher the estimated reliability [19]. In this study, the distribution of data has spread from score 1 to score 14 from a total score of 21, but it is not evenly distributed following a normal curve.

Tests that are too easy or too difficult tend to give low reliability scores [19]. This is because easy and difficult tests produce a limited distribution of scores. In addition, the ideal average difficulty desired for multiple-choice tests can be estimated by taking the midpoint between the expected chance score and the maximum possible score. The expected chance score for multiple-choice questions with four answer choices is 25% correct (one out of four). This means that the ideal average difficulty of the test should be 62.5% (0.625). While the average difficulty in this study was 0.329 which is less than 0.625. This is one of the reasons that the reliability of the test is not so high.

Based on the results of the study above, it can be concluded that the HOTS-characterized mathematics learning outcome test in the form of multiple choices for fifth grade elementary school students has 15 good questions that meet both qualitative and quantitative criteria. However, the fifteen questions have a reliability that is not so high. Even so, the fifteen questions can be an example to measure and develop the HOTS of fifth grade elementary school students in mathematics.

Furthermore, although the developed HOTS-characterized mathematics learning outcome test consists of 15 items that meet both qualitative and quantitative criteria, a comparison with international standardized tests such as PISA is necessary. The PISA assessment focuses on mathematical literacy and real-world problem-solving, which aligns with the principles of HOTS. Therefore, future studies are encouraged to align or benchmark the developed test items against PISA frameworks to ensure global relevance and enhance the validity of HOTS assessments at the elementary level.

## 4. CONCLUSION

Based on the results and discussion, it can be concluded that out of 21 questions created, there are 15 questions that can be used to measure HOTS of Elementary School students in mathematics, while 6 other questions were declared to be failed. 1 question was failed during the validity test and 5 questions were failed during the analysis of the question difficulty index. Of the 15 questions retained, 14 questions were categorized as difficult and 1 question with a moderate level of difficulty. Judging from the discriminating power of the questions, tThere are 2 items with a high category, 9 items with a medium category, and 4 items with a low category. All distractors in each item function effectively. The results of the reliability test conducted on 15 items that were still maintained showed thatThe fifteen questions have a reliability that is not so high, which is 0.609. However, the fifteen questions can be an example to measure and develop HOTS of grade V elementary school students in mathematics.

The suggestions that can be put forward in this study are as follows: 1) for students, it is necessary to get used to working on HOTS mathematics problems in order to have high abilities in mathematics learning outcome tests; 2) for teachers, these HOTS problems can be used together with the implementation of mathematics learning strategies such as problem-solving methods or problem-based learning (PBL) models; teachers are expected to use contextual problem-solving problems and not routine ones; 3) for schools, it is expected to be able to accommodate and encourage the creativity of teachers and students in all forms of learning, especially in mathematics learning well; 4) for other researchers, the results of this analysis can be used as an example that HOTS problems can be developed in elementary schools, especially in the field of elementary school mathematics.

**REFERENCES**

[1] A. H. Abdullah, N. L. Z. Abidin, & , M. Ali, “Analysis of students’ errors in solving Higher Order Thinking Skills (HOTS) problems for the topic of fraction”, *Asian Social Science*, vol. 11, no. 21, pp. 133-142, 2015. Available on link https://epe.bac-lac.gc.ca/100/201/300/asian_social_science/2015/ASS-V11N21-All.pdf?nodisclaimer=1

[2] L. W. Anderson, & D. R. Krathwohl (Eds), “Kerangka Landasan untuk Pembelajaran Pengajaran, dan Asesmen. Translated by Agung Prihantoro, “*A Taxonomy for Learning, Teaching, and Assessing: A Revision of Taxonomy of Educational Objectives*”, Yogyakarta, Pustaka Belajar., 2015.

[3] E. Apino, & H. Retnawati, “Developing Instructional Design to Improve Mathematical Higher Order Thinking Skills of Students”. *Journal of Physiscs: Conference Series*, vol. 812, no. 1, pp. 1-7, 2017. DOI: https://doi.org/10.1088/1742-6596/812/1/012100

[4] K. Arafah, B. D. Amin, S. S. Sari, & A. Hakim, “The development of higher order-thinking skills (HOTS) instrument assessment in physics study”, *Journal of Physics: Conference Series*, vol. 1899, No. 1, 2021. IOP Publishing. DOI: https://doi.org/10.1088/1742-6596/1899/1/012140

[5] I. K. S. Ariyana, “Analisis Kebutuhan Tes Hasil Belajar Matematika Sekolah Dasar Berkarakteristik Hots Berbentuk Pilihan Ganda”, *Jurnal Pendidikan dan Kebudayaan Missio*, vol. 12, no. 2, pp. 87-98, 2020. Available on link https://scholar.archive.org/work/s5v7t4dcnzeozdd5pgbrgu3f5y/access/wayback/http://unikastpaulus.ac.id/jurnal/index.php/jpkm/article/download/453/353

[6] J. Astutik, “Developing Save Your Food Kit (Sayofu Kit) to Support Inquiry, Improve Student Learning Outcomes at SMP Plus Hidayatul Mubtadiin and Public Awareness on Food Additives”. *Journal of Physiscs: Conference Series*, vol. 812, no. 1, pp. 1-8, 2017. DOI: https://doi.org/10.1088/1742-6596/812/1/012080

[7] S. M. Brookhart, “*How to Assess Higher-Order Thinking Skills in Your Classroom*”, Virginia USA, ASCD, 2010. Available on link https://www.daneshnamehicsa.ir/userfiles/files/1/7-%20How%20to%20Assess%20Higher-Order%20Thinking%20Skills%20in%20Your%20Classroom%20-%20Copy%201.pdf

[8] M. I. E. Budiarti, F. Musa’ad, & S. Supriadi, “Analisis Proses Pemecahan Masalah Peserta Didik Dalam Menyelesaikan Soal Berorientasi PISA”. *Prima Magistra: Jurnal Ilmiah Kependidikan*, vol. , no. 4, pp. 533-539, 2023. DOI: https://doi.org/10.37478/jpm.v4i4.3051

[9] I M. Candiasa, “*Pengujian Instrumen Penelitian Disertai Aplikasi ITEMAN dan BIGSTEPS*”, Singaraja, Unit Penerbitan UNDIKSHA, 2010.

[10] N. Dantes, “*Metode Penelitian*”, Yogyakarta, ANDI, 2012.

[11] D. DiBattista, & L. Kurzawa, “Examination of the Quality of Multiple-choice Items on Classroom Tests”. *The Canadian Journal for the Scholarship of Teaching and Learning*, vol. 2, no. 2, pp. 1-23, 2011. Available on link https://eric.ed.gov/?id=EJ985723

[12] F. I. El-Uri, & N. Malas, “Analysis of Use of A Single Best Answer Format in an Undergraduate Medical Examination”. *Qatar Medical Journal*, vol. 2013 no. 1, pp. 3-6, 2013. DOI: https://doi.org/10.5339/qmj.2013.1

[13] L. U. Fatimah, & K. Alfath, “Analisis kesukaran soal, daya pembeda dan fungsi distraktor”. *AL-MANAR: Jurnal Komunikasi dan Pendidikan Islam*, *8*(2), 37-64, 2019. DOI: https://doi.org/10.36668/jal.v8i2.115

[14] F. M. Firdaus, “Pembelajaran Matematika Realistik Untuk Meningkatkan Kemampuan Pemecahan Masalah Matematis Siswa Sekolah Dasar (Penelitian Tindakan Kelas dalam Pembelajaran Pecahan di Kelas V SDN Nangela Kabupaten Bandung)”, *PEDAGOGIK*, vol. 3, no. 1, pp. 80-88, 2015. DOI: https://doi.org/10.33558/pedagogik.v3i1.1262

[15] M. Firja MKS, “Metode Pembelajaran Matematika yang Efektif dalam Meningkatkan High Order Thinking Siswa”, *Journal on Education*, vol. 6, no. 4, pp. 22406-22415, E-ISSN: 2654-5497, P-ISSN: 2655-1365, 2024. DOI: https://doi.org/10.31004/joe.v6i4.6407

[16] K. Fitriani, & Maulana. “Meningkatkan Kemampuan Pemahaman Dan Pemecahan Masalah Matematis Siswa SD Kelas V Melalui Pendekatan Matematika Realistik”, *Mimbar Sekolah Dasar*, vol. 3, no. 1, pp. 40-52, 2016. DOI: https://doi.org/10.53400/mimbar-sd.v3i1.2355

[17] S. Giarti, “Peningkatan Keterampilan Proses Pemecahan Masalah Dan Hasil Belajar Matematika Menggunakan Model Pbl Terintegrasi Penilaian Autentik Pada Siswa Kelas VI SDN 2 Bengle, Wonosegoro”, *Scholaria: Jurnal Pendidikan Dan Kebudayaan*, vol. 4, no. 3, 13-27, 2014. DOI: https://doi.org/10.24246/j.scholaria.2014.v4.i3.p13-27

[18] G. E., Gilbert, & S. Prion, “Making Sense of Methods and Measurments: Lawshe’s Content Validity Index”. *Clinical Simulation in Nursing*, vol. 12, no. 12, pp. 530-531, 2016. DOI: https://doi.org/10.1016/j.ecns.2016.08.002

[19] N. E. Gronlund, “*Measurement and Evaluation in Teaching*”, New York, Macmillan Publishing, 1981.

[20] A. Gymnastiar, N. N. A’fiani, N. B. Ramadhan, & A. N. Aeni, “Pengembangan Media Articulate Storyline Untuk Pembentukan Akhlak Terpuji Siswa Kelas VI SD”, *Al Qalam: Jurnal Ilmiah Keagamaan Dan Kemasyarakatan*, *16*(4), 1241-1248, 2022. DOI: http://dx.doi.org/10.35931/aq.v16i4.1062

[21] Lutfianto, et al., “Unfinished Student Answer in PISA Mathematics Contextual Problem”. *IndoMS-JME*, vol. 4, no. 2, pp. 188-193, 2013. Available on link https://eric.ed.gov/?id=EJ1078951

[22] A.R. Menon, & P. N. Kannambra, “Item Analysis to Identify Quality Multiple Choice Questions”. *National Journal of Laboratory Medicine*, vol. 6, no. 2, pp. 7-10, 2017. Available on link https://njlm.net/articles/PDF/2215/25690_CE[VSU]_F(GH)_PF1(VsuGH)_PFA(GH)_PF2(VsuGH).pdf

[23] NCTM (National Council Teacher of Mathematics), “Principles and Standard for School Mathematics”, Restin VA, NCTM, 2000.

[24] N. R. Nengsih, E. Yusmaita, & F. Gazali, “Evaluasi validitas konten dan konstruk bahan ajar asam basa berbasis REACT”, *EduKimia*, *1*(1), 1-10, 2019. https://www.researchgate.net/publication/335963036_Evaluasi_Validitas_Konten_dan_Konstruk_Bahan_Ajar_Asam_Basa_Berbasis_REACT

[25] P. P. Noor, & A. P. Abadi, "Kemampuan berpikir tingkat tinggi dalam perkembangan pembelajaran matematika SMA". *Jurnal Educatio Fkip Unma*, vol. 8, no. 2, pp. 466-473, 2022. DOI: https://doi.org/10.31949/educatio.v8i2.1986
[26] J. C. Nunnally, *Psychometric Theory*, New York, McGraw-Hill, 1978.
[27] S.C. Ohiri, & R. O. Okoye, "Application of classical test theory as linear modeling to test item development and analysis", *International Research Journal of Modernization in Engineering Technology and Science*, *5*(1), 2152-2159, 2023. https://www.researchgate.net/publication/379511621_APPLICATION_OF_CLASSICAL_TEST_THEORY_AS_LINEAR_MODELING_TO_TEST_ITEM_DEVELOPMENT_AND_ANALYSIS
[28] Partnership for 21st Century Skills, "*Framework for 21st Century Learning*", 2006. Available on link https://www.marietta.edu/sites/default/files/documents/21st_century_skills_standards_book_2.pdf.
[29] Partnership for 21st Century Skills, "*21st Century Skills, Education & Competitiveness: A Resource and Policy Guide*", 2008. Available on link https://files.eric.ed.gov/fulltext/ED519337.pdf.
[30] Permendikbud RI Nomor 67 Tahun 2013 tentang Kurikulum SD.
[31] E. Purwaaktari, "Pengaruh Model Collaborative Learning terhadap Kemampuan Pemecahan Masalah Matematika dan Sikap Sosial Siswa Kelas V SD Jarakan Sewon Bantul". *Jurnal Penelitian Ilmu Pendidikan*, vol. 8, no. 1, pp. 95-111, 2015. DOI: https://doi.org/10.21831/jpipfip.v8i1.4932
[32] Y. Y. Putra, et al., "Pengembangan Soal Matematika Model PISA Konten Bilangan Untuk Mengetahui Kemampuan Literasi Matematika Siswa". *Jurnal Elemen*, vol. 2, no. 1, pp. 14-26, 2016. DOI: https://doi.org/10.29408/jel.v2i1.175
[33] Richey, R. C., & Klein, J. D., "*Design and development research: Methods, strategies, and issues*", New Jersey, Routledge, 2007.
[34] J. Rubin, & M. Rajakaruna, "Teaching and Assessing Higher Order Thinking in the Mathematics Classroom with Clickers". *iSER, Mathematics Education*, vol. 10, no. 1, pp. 37-51, ISSN: 1306-3030, 2015. DOI: https://doi.org/10.29333/iejme/290
[35] Ruhyana, "Analisis Kesulitan Siswa dalam Pemecahan Masalah Matematika". *Jurnal Computech & Bisnis*, vol. 10, no. 2, pp. 106-118, 2016. Available on link https://download.garuda.kemdikbud.go.id/article.php?article=568508&val=7183&title=Analisis%20Kesulitan%20Siswa%20Dalam%20Pemecahan%20Masalah%20Matematika
[36] C. Santi, A. Helmon, & E. Sennen, "Pembelajaran Pemecahan Masalah Matematika Di Sekolah Dasar: Masalah Dan Upaya Yang Dapat Dilakukan". *Jurnal Literasi Pendidikan Dasar*, vol. 2, no. 2, pp. 31-40, 2021. Available on link https://www.neliti.com/publications/408790/pembelajaran-pemecahan-masalah-matematika-di-sekolah-dasar-masalah-dan-upaya-yan
[37] C. L. Scott, "The Futures of Learning 2: What Kind of Learning for the 21st Century?", *UNESCO Education Research and Foresight*, *ERF Working Papers Series*, no. 14, 2015. Available on link https://repositorio.minedu.gob.pe/handle/20.500.12799/3709
[38] D. Scully, "Constructing Multiple-Choice Items to Measure Higher-Order Thinking", *Practical Assessment, Research, and Evaluation*, vol. 22, no. 1, 2017. DOI: https://doi.org/10.7275/swgt-rj52
[39] A. N. Shete, & A. Kausar, "Item analysis: An evaluation of multiple choice questions in Physiology examination". *Journal of Contemporary Medical Education*, vol. 3, no. 3, pp. 106-109, 2015. Available on link https://www.academia.edu/download/82874839/item-analysis-an-evaluation-of-multiple-choice-questions-in-physiology-examination.pdf
[40] K. S. Shultz, et al., "*Measurement Theory in Action Case Studies and Exercises*", New York, Routledge, 2014.
[41] T. Sonnabend, "*Mathematics for Teachers: An Interactive Approach for Grades K–8 Fourth Edition*", USA, Brooks/Cole, 2010.
[42] I G. P. Suharta, "Kemampuan Siswa Sekolah Dasar Dalam Pemecahan Masalah Matematika Real". *Jurnal Pendidikan dan Pengajaran*, vol. 49, no. 3, pp. 137-147, 2016. DOI: https://doi.org/10.23887/jppundiksha.v49i3.9018
[43] R. Sumarni, "Kemampuan Guru Menyusun Soal HOTS dalam Bentuk Pilihan Ganda di SDN 02 Legokgunung". *Action Research Journal*, vol. 1, no. 3, pp. 262-265, 2022. DOI: https://doi.org/10.51651/arj.v1i3.77
[44] N. M. Tajudin, Mathematical Knowledge and Higher Order Thinking Skills For Teaching Algebraic Problem Solving, *SOCIOINT15, Proceedings of 2nd International Conference on Education, Social Sciences and Humanities*, pp. 8-10, 2015, Available on link http://www.ocerint.org/socioint15_epublication/papers/54.pdf
[45] Usmaedi, "Menggagas Pembelajaran HOTS pada Anak Usia Sekolah Dasar", *JPSD*, vol. 3, no. 1, pp. 82-95, ISSN 2540-9093, 2017. DOI: http://dx.doi.org/10.30870/jpsd.v3i1.1040
[46] J. A.Van de Walle, et al., "*Elementary and middle school mathematics teaching developmentally (8th)*", United States of America, Pearson Education, 2013.
[47] S. Wardhani, et al., "*Modul Matematika SD Program BERMUTU: Pembelajaran Kemampuan Pemecahan Masalah Matematika di SD*", Yogyakarta, PPPPTK Matematika, 2010.
[48] I W. Widana, "Higher Order Thinking Skills Assessment (HOTS)". *JISAE*, vol. 3, no. 1, pp. 32-44, 2017. Available on link https://www.academia.edu/download/73441255/c61336810e46336d43c29c4195c8d86dfd66.pdf
[49] A. Wijaya, "The Relationships between Indonesian Fourth Graders' Difficulties in Fractions and the Opportunity to Learn Fractions: A Snapshot of TIMSS Results". *International Journal of Instruction*, vol. 10, no. 4, pp. 221-236, 2017. Available on link https://eric.ed.gov/?id=EJ1163264
[50] A. Wijaya, et al., "Opportunity-to-Learn Context-Based Tasks Provided by Mathematics Textbooks". *Educational Studies in Mathematics*, vol. 89, no. 1, pp. 41-65, 2015. DOI: https://doi.org/10.1007/s10649-015-9595-1
[51] A. A. Zahra, & A. K. At-Taqiyyah, "Mengukur kemampuan kognitif tingkat tinggi dengan pertanyaan pilihan ganda". *Jurnal Pelita Nusantara*, vol. 1, no. 4, pp. 462-470, 2024. DOI: https://doi.org/10.59996/jurnalpelitanusantara.v1i4.336