e-ISSN: 2721-1274
p-ISSN: 2721-2718

THE DEVELOPMENT OF TEST INSTRUMENT BASED ON HIGHER ORDER THINKING SKILLS (HOTS) WITH ADDIE MODEL

Silvania Carella Prinst[1], Rita Juliani[2]
[1,2]Department of Physics, State University of Medan, Indonesia
E-mail: carellasilvania@gmail.com

ABSTRACT
This study aims to develop a physics test instrument based on Higher Order Thinking Skills (HOTS) with ADDIE Model that meets the qualification in the aspects of validity, reliability, discriminating power, level of difficulty, and effectiveness of distractors. The type of research used is the Research and Development (R&D) ADDIE model with 5 stages, namely: (1) the analysis: problem identification, need analysis, and test instrument analysis, (2) design: format selection and initial design of HOTS test instruments, (3) development: HOTS test instrument design development, content validation by experts, revision of the results of content validation, (4) implementation: small class tests on 10 students of class XI IPA 2, large class tests on 45 students of class XI IPA 2 and XI IPA 3 in SMA Swasta PAB-8 Saentis, (5) evaluation: re-analyzing the results of the trial, and conclusions from all stages of the research. The test instrument developed consisted of 20 multiple-choice items. The results of the content validation test obtained valid test instruments with revisions to material, construction, and language aspects. The results of the small class test showed that 85% of valid items, very reliable, 95% of items had good discriminating power, 95% had moderate difficulty, and 70% had good distractors. Large class test results obtained 88% of items are valid, reliable, 56% of items have good discriminating power, 82% of items have moderate difficulty, and 82% of items have good distractor effectiveness, so the test instrument is feasible to use to measure students higher order thinking skills.

Keywords: Higher Order Thinking Skills, Test Instrument, ADDIE Model

INTRODUCTION

Education is a conscious effort carried out in a structured manner in the learning process to gain knowledge, skills and insights in order to develop self potential. Education can be said to be successful if the quality of education can be achieved properly, can create quality human resources (HR) that have the potential, and can deal with challenges that will come in the future. In the current era of globalization, the quality of education has become of concern from various circles.

The quality of education in Indonesian at the global level has not shown satisfactory results. Indonesian is ranked 69th out of 76 countries in the PISA (Program for International Student Assessment) results reported by the Organization for Economic Co-Operation and Development (OECD, 2012). Education faces very complex problems, where in the era of the industrial revolution 4.0 emphasizing the digital economy, artificial intelligence, big data, and robotics, demanding that the world of education constructs creativity, critical thinking, mastery of technology, and digital literacy skills (Wahyuni, 2018).

e-ISSN: 2721-1274
p-ISSN: 2721-2718

Students are required to be capable of Higher Order Thinking Skills called HOTS. The main purpose of HOTS is how to improve students higher order thinking skills, expecially in critical thinking skills to receive various types of information, think critically in solving a problem using knowledge possessed, able to make conclusions, and decisions in complex situations (Saputra, 2016). Higher order thinking skills can be accommodate through the curriculum contained in education which is used as a guide in the implementation of learning.

Assessment is needed to measure students higher order thinking skills. Assessment is an evaluation of student achievement in learning to see students learning outcomes carried out by educators. The instrument in cognitive assessment is a test. Test are used to measure learning. The test developed is the HOTS test which involves higher order thinking, cognitive and complex problems.

The development of HOTS based test instruments is the creation of test instruments or question based on HOTS standards. With HOTS students are not only limited to memorizing or understanding existing concepts, but are able to analyze the state of the information obtained, be able to relate one information to another, solve existing problems, and be able to create new ideas. The problem that is often faced in schools is that the questions given to students tend to test aspects of memory such as questions that tend to be about memorizing formulas, the lack of questions that test students thinking, and analytical skills in solving problems.

Among the factors causing the low thinking ability of students are the lack of training of students in solving HOTS based questions, the problem faced by teachers are the inability of teachers to develop HOTS based test instruments, and the unavailability of instruments specially designed to train students abilities in solving HOTS based questions. Based on the identification of the above problems, it is necessary to have a test instrument that can build students science literacy, and improve students higher order thinking skills. Thus, with the developed test instrument, teacher can train students ability to solve HOTS based questions.

According to Dewanto in Purbaningrum (2017), "Higher order thinking is a capacity above the information given, with a critical attitude to evaluate, have metacognitive awareness, and have problem solving abilities". Higher order thinking skills are ability to understand, analyze, and solve problems in various ways from different perspectives according to students abilities. In other words, higher order thinking is not just memorizing and conveying known information, but is able to connect, manipulate, transform knowledge, experience in an effort to make decisions, and solve problems in new situations.

The realm of taxonomy bloom is used to measure higher order thinking skills, and indicators to measure higher order thinking skills include analyzing (C4), evaluating (C5), and creating (C6) as revealed by Krathwohl (Ayuningtyas & Rahaju, 2012). Karthwohl explains the indicators as follows: (1) analyzing (C4) is the process of separating material into its constituent parts, and detecting the relationship between one part and another, (2) evaluate (C5) is making decisions such as checking based on criteria and standards, (3) create (C6) is to form a coherent whole by putting elements together, or creating an original result, such as compiling, planning, ang generating. There are several aspects that show a persons higher order thinking skills, namely critical thinking skills, creative thinking skills, and problem solving (Dewi *et al.*, 2015).

Critical and creative thinking skills are indicators of higher order thinking. Critical and creative thinking skills in students can be measure by developing a test

e-ISSN: 2721-1274
p-ISSN: 2721-2718

instrument based on Higher Order Thinking Skills. Based on this description, it is considered necessary to conduct further research. Therefore, a research entitled: "The Development of Test Instrument Based on Higher Order Thinking Skills (HOTS) with ADDIE Model".

RESEARCH METHODS

This type of research is Research and Development (R&D). Research and Development is a research method that aims to produce certain products which are then tested for validity and effectiveness in implementing these products (Hanafi, 2017). This research was conducted at SMA Swasta PAB-8 Saentis which is located at Jalan Kali Serayu Dusun 16, Saentis, Percut Sei Tuan district, Deli Serdang Regency, North Sumatera 20371.

This research will produce a product in the form of items that can later be used as a test instrument in physics subjects, especially on translational dynamics. The type of data used in this research are qualitative and quantitative data. Qualitative data is generated from the review of items by experts consisting of 3 experts on the test instrument. While quantitative data is generated from the results of the development test instrument based on HOTS the material of Translational Dynamics to measure validity, reliability, discriminating power, level of difficulty, and effectiveness of distractors.

The population of this research is all students of class XI IPA 2 and XI IPA 3 SMA Swasta PAB-8 Saentis. The research sample for the small class is 10 students class XI IPA 2 and the large class is 45 students class XI IPA 2 and XI IPA 3 which have similarities in the learning time of the material of translational dynamics. This research used ADDIE model. The ADDIE model is a research model that produces a particular product and then tests the effectiveness of that product (Hartini & Martin, 2020). The five stages of ADDIE when presented in chart form are as follows:

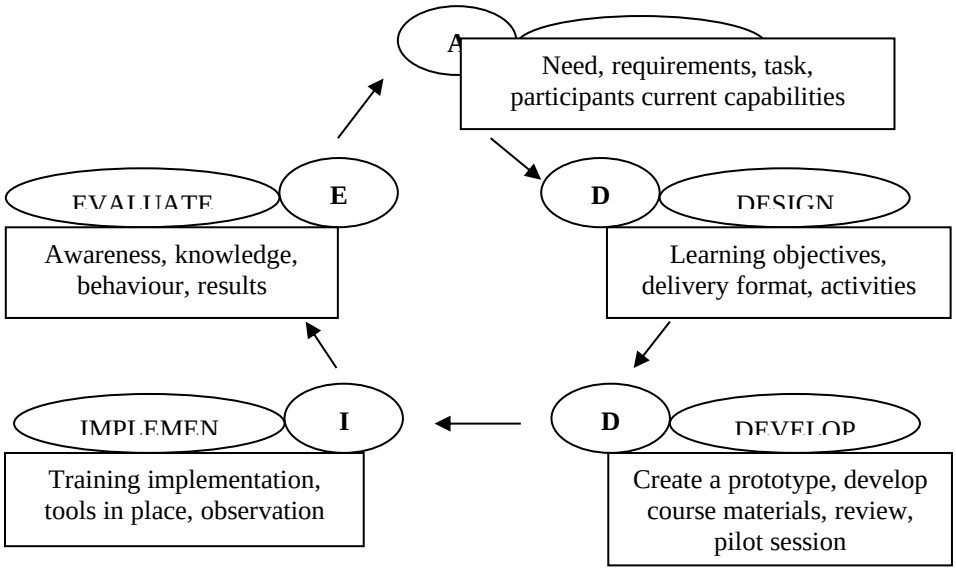

Figure 1. Stages for ADDIE model (Sugiyono, 2015)

The stages of this research are shown in the scheme below:

e-ISSN: 2721-1274
p-ISSN: 2721-2718

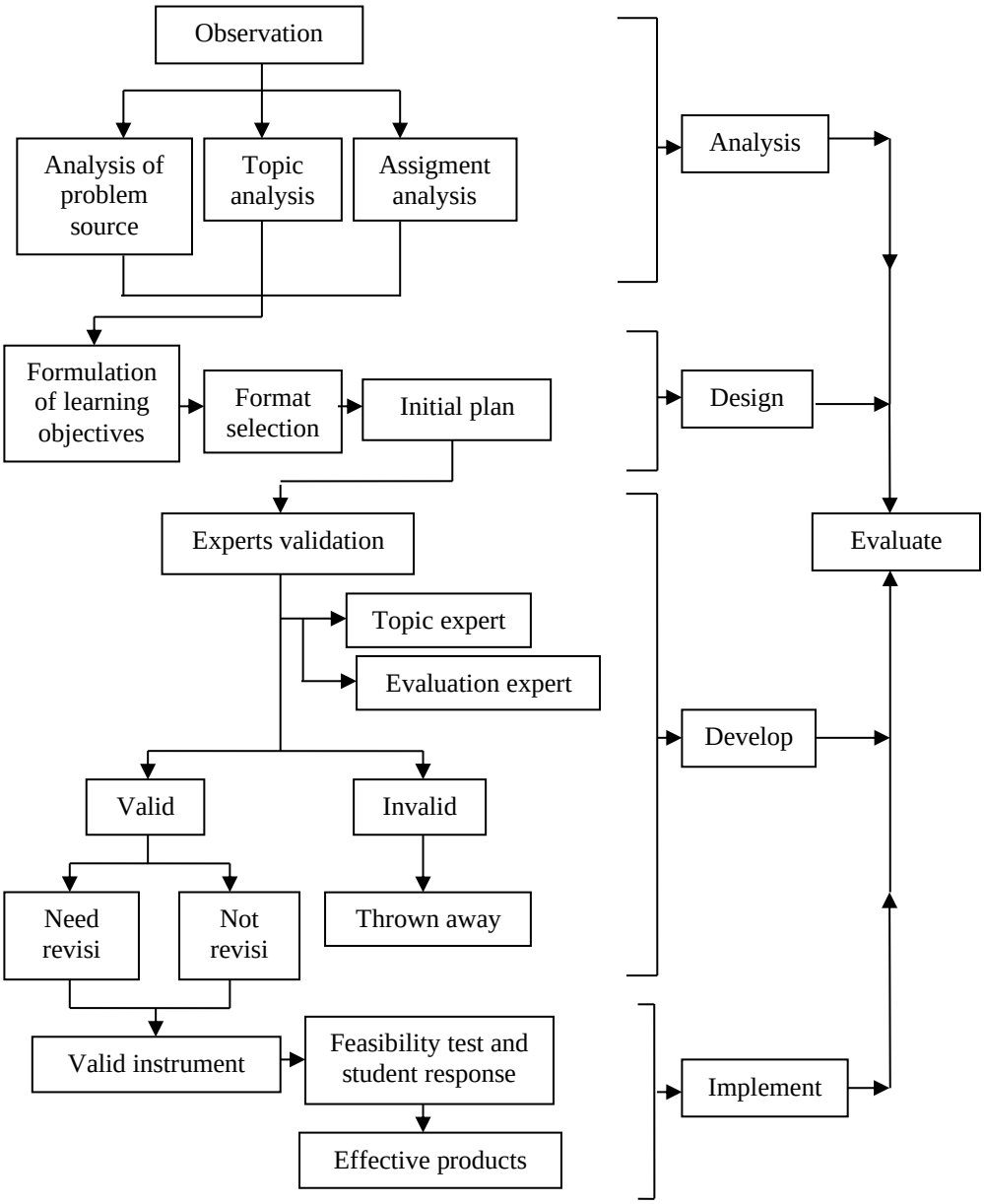

Figure 2. Research Flowchart

RESULTS AND DISCUSSION

A. Research Result

1. Analysis Stage

Based on the results of interview with SMA Swasta PAB-8 Saentis teachers, information was obtained that the majority of students scientific literacy in physics learning was still categorized as low. This is in accordance with the 2018 PISA results, which show that Indonesian is ranked 64th out of 72 countries whose scientific literacy is still relatively low. Students interest in learning is very low, and free hours are filled with playing with gadgets.

Teachers have not implemented HOTS questions because students have not been able to accept, and have not been motivated to learn physics. The questions used by teachers still tend to test the memory aspect because calculation questions are still something that students are afraid of. Teachers often hold quizzes before learning

e-ISSN: 2721-1274
p-ISSN: 2721-2718

begins to motivate students to study physics. From the analysis stage, it is known that students higher order thinking skills are still very low because the test instruments used in school still test aspects memory, so students are not yet familiar with HOTS based question. So it is necessary to develop HOTS based test instruments.

2. Design Stage

The format used in instrument development is a test based on higher order thinking skills (HOTS). It consists of 20 HOTS questions with 5 answer choice (A, B, C, D, and E). The material used is Translational Dynamics. The components contained in the HOTS based test instrument grid consist of material, basic competencies, competency achievement indicators, class/semester, question indicators, cognitive level, question form, and question number. Cognitive level is determined for each test item based on indicators. The cognitive levels used are C4, C4, and C6.

Next is preparing the question stimulus. The question stimulus must be able to encourage students to read the stimulus, meaning the stimulus used must be interesting. Generally, interesting stimuli are ones that students have never read, are new, or are issues that are currently emerging. Meanwhile, contextual stimulus is a stimulus that is in accordance with the realities of everyday life and encourages students to read.

Instructions for working on questions are instructions that students can follow in answering the questions provided. And the question card consists of subject, class/semester, curriculum, basic competencies, competency achievement indicators, material, question indicators, cognitive level, question description, multiple choice, answer key, and HOTS question characteristics for each question item. Test students who answer test items correctly get 1 point, and students who answer incorrectly get 0 point.

3. Development Stage

The initial design that has been created is then developed further until it becomes a product that can be validated by experts. A summary of these products can be seen in table 1:

Table 1. Summary of Based Test Products (HOTS)

Developed Products	Content
Test higher order thinking skills	Material, basic competencies, competency achievement indicators, class/semester, question indicators, cognitive level, question form, and question number
Instructions for taking HOTS based tests	Choose the most appropriate answer by circling the correct answer
Question sheet for HOTS based tests	20 HOTS based questions
Answer key	Answer each test item from the 20 questions
Scoring quidelines	A score pf is 1 given for the right answer, and a score of 0 is given for the wrong answer

The test instrument was developed from 15 multiple-choice items to 20 multiple-choice items. Before being tested on students, the HOTS based test instrument is validated by experts. The type of validity used is content validity, which includes three aspects of material, construction, and language assessment.

e-ISSN: 2721-1274
p-ISSN: 2721-2718

The test instrument was assessed by experts using a Likert scale with 4 levels, namely 1, 2, 3,and 4, with the provisions 1 = invalid, 2 = sufficient, 3 = valid, and 4 = very valid. Quantitative content validation data is processed using the test item validation index equation proposed by Aiken (Index V). The V Index value ranges between 0-1. The data is processed using the Microsoft Excel. The results of the content validation test on the test instrument obtained an average validation value of 0.846 valid with revisions. The results of the content validation test are listed in table 2 below:

Table 2. Result of Material, Construction, and Language Content Validation Test

Question Number	Validate Aspect Content			Validation Per Item
	Content of Material	Construction	Language	
1	0.73	0.71	0.81	0.75
2	0.69	0.76	0.82	0.76
3	0.76	0.78	0.82	0.79
4	0.78	0.93	0.87	0.86
5	0.74	0.73	0.93	0.80
6	0.88	0.82	0.87	086
7	0.78	0.92	0.98	0.89
8	0.76	0.72	0.83	0.78
9	0.81	0.86	0.97	0.88
10	0.92	0.96	0.92	0.93
11	0.83	0.84	0.98	0.89
12	0.78	0.79	1.00	0.86
13	0.77	0.90	0.96	0.87
14	0.62	0.71	1.00	0.78
15	0.77	0.79	1.00	0.85
16	0.91	0.89	1.00	0.93
17	0.73	0.70	1.00	0.81
18	0.90	0.82	1.00	0.91
19	0.64	0.70	1.00	0.78
20	0.91	0.91	1.00	0.94

4. Implementation Stage

After the questions have been validated by experts, revisions have been made to the questions that need to be revised, and input from experts has been added to the questions. Next, researchers carried out the implementation stage. The implementation stage is carried out in small and large class to determine the validity, reliability, level of difficulty, discriminating power, and effectiveness of distractors.

a. Small Class Trial

The small class test was carried out at SMA Swasta PAB-8 Saentis Percut Sei Tuan class XI IPA 2 with a sample of 10 students. Students answer question within 60 minutes. Data analysis in this research was carried out using Microsoft Excel.

a.1. Validity

The results of the content validity test are quantitative data with an average validation value of 0.623 and the data shown that there are 17 valid items, and 3 invalid items (number 3, 6, and 18). The data was analyzed using a rough numerical product moment correlation equation. The V index value ranges between 0-1 with a significance level of 5%. A test item is said to be valid if $r_{XY\ count} > r_{XY\ table}$. The recapitulation results of the validity test items in small classes are shown in table 3:

e-ISSN: 2721-1274
p-ISSN: 2721-2718

Table 3. Result of the Small Class Validity Test

Question Number	Validity		Description
	R_{count}	R_{table}	
1	0.6347	0.4973	Valid
2	0.6001		Valid
3	0.4092		Invalid
4	0.798		Valid
5	0.9354		Valid
6	0.3205		Invalid
7	0.6488		Valid
8	0.7706		Valid
9	0.7363		Valid
10	0.7363		Valid
11	0.6274		Valid
12	0.6274		Valid
13	0.5011		Valid
14	0.6683		Valid
15	0.7217		Valid
16	0.55918		Valid
17	0.6956		Valid
18	0.3718		Invalid
19	0.5319		Valid
20	0.566		Valid

a.2. Reliability

The reliability of the instrument was analyzed using the Cronbach's Alpha equation. The data shown is very high reliability because $r \geq 0.80$ and the result of reliability test instrument of small class can be seen in table 4 below:

Table 4. Result of Reliability Test Instrument of Small Class

No	Number of Items	Reliability	Description
1	20	0.921	Very High Reliability

a.3. Discriminating Power

The results of the discriminating power tests on the test items are listed briefly in table 5 below:

Table 5. Result of Discriminating Power Test Instrument of Small Class

No	Discriminating Power Coefficient	Discriminating Power Criteria	Question Number
1.	0.71 ≤ DP ≤ 1.00	Very Good	4, 5, 8, 14, 17
2.	0.41 ≤ DP ≤ 0.70	Good	1, 2, 3, 6, 7, 9, 10, 11, 12, 13, 15, 16, 19, 20
3.	0.21 ≤ DP ≤ 0.40	Enough	18
4.	0.00 ≤ DP ≤ 0.20	Not Good	0
5.	0 < DP	Very Not Good	0

a.4. Level of Difficulty

Difficulty analysis is needed to identify items in the easy, medium, and difficult categories. In the results of the analysis of the 20 items tested, there were 19 items classified as medium (number 1, 2, 4, to 20) and 1 item (number 3) classified as difficult. The level of difficulty of the test instrument is in the medium category at 80% and the difficult category at 20%.

a.5. Effectiveness of Distractors

e-ISSN: 2721-1274
p-ISSN: 2721-2718

Items are considered good if they have distractors that effectively deceive students. The distractors functions well if it has been selected by more than 5% of test participants (Sudijono, 2008). The effectiveness of the test item distractors is shown in table 6 below:

Table 6. Result of the Effectiveness of Distractors Test Instrument in Small Class

Question Number	Effectiveness of Distractors Percentage					Effective Number of Distractors	Description
	A	B	C	D	E		
1	0	20	50*	20	10	3	Good
2	0	20	20	10	50*	3	Good
3	20	20*	0	30	30	3	Good
4	60*	0	40	0	0	2	Enough
5	0	10	30	50*	10	3	Good
6	30	0	10	60*	0	2	Enough
7	0	10	20	0	70*	3	Good
8	0	40*	0	60	0	1	Not Good
9	30*	20	40	10	0	3	Good
10	30	30	30*	10	0	3	Good
11	20	10	10	60*	0	3	Good
12	30	10	0	0	60*	2	Enough
13	10	50*	40	0	0	2	Enough
14	0	0	40*	40	20	3	Good
15	70*	0	0	0	30	2	Enough
16	0	60*	10	30	0	3	Good
17	0	0	30	10	60*	3	Good
18	10	10	30*	40	10	3	Good
19	40*	20	0	20	20	3	Good
20	0	40*	50	0	10	3	Good

Description: *Answer Key

b. Large Class Trial

The large class test was carried out at SMA Swasta PAB-8 Saentis Percut Sei Tuan, involving 45 students from classes XI IPA 2 AND XI IPA 3 in 60 minutes.

b.1. Validity

The recapitulation results of the validity test items in large classes are shown in table 7:

Table 7. Result of the Large Class Validity Test

Question Number	Validity		Description
	R_{count}	R_{table}	
1	0.742	0.242	Valid
2	0.570		Valid
4	0.256		Valid
5	0.508		Valid
7	0.646		Valid
8	0.219		Invalid
9	0.590		Valid
10	0.559		Valid
11	0.431		Valid
12	0.696		Valid
13	0.517		Valid

e-ISSN: 2721-1274
p-ISSN: 2721-2718

14	0.513		Valid
15	0.570		Valid
16	0.238		Invalid
17	0.546		Valid
19	0.592		Valid
20	0.250		Valid

b.2. Reliability

The reliability of the instrument was analyzed using the Cronbach's Alpha equation. The calculation result of the large class reliability test is 0.808 has high reliability. The reliability result of the large class test can be seen in table 8 below:

Table 8. Result of Reliability Test Instrument of Large Class

No	Number of Items	Reliability	Description
1	20	0.830	High Reliability

b.3. Discriminating Power

The discriminating power of test items is needed to differentiate intelligence between test takers. The differentiating power in research test instrument is 9 items in the good category, 6 items in the fair category, and 2 items in the not good category. The results of the discriminating power tests are listed briefly in table 9 below:

Table 9. Result of Discriminating Power Test Instrument of Large Class

No	Discriminating Power Coefficient	Discriminating Power Criteria	Question Number
1.	0.71 ≤ DP ≤ 1.00	Very Good	-
2.	0.41 ≤ DP ≤ 0.70	Good	1, 2, 7, 9, 10, 12, 15, 17, 19
3.	0.21 ≤ DP ≤ 0.40	Enough	4, 5, 11, 13, 14, 20
4.	0.00 ≤ DP ≤ 0.20	Not Good	8, 16
5.	0 < DP	Very Not Good	0

b.4. Level of Diffuculty

An analysis of the difficulty level of test items is needed to find items in the easy, medium, and difficult categories. The results of the analysis level of difficulty test items showed that 3 out of 17 items were at a difficult level, and 14 out of 17 items were at a medium level. Level of difficulty of the test instrument is in the medium category at 82% and the difficult category at 18%.

b.5. Effectiveness of Distractors

Items are considered good if they have distractors that effectively deceive students. The distractors functions well if it has been selected by more than 5% of test participants (Sudijono, 2008). The effectiveness of the test item distractors is shown in table 10 below:

Table 10. Result of the Effectiveness of Distractors Test Instrument in Large Class

Question Number	Effectiveness of Distractors Percentage					Effective Number of Distractors	Description
	A	B	C	D	E		
1	4	24	49*	4	18	4	Very Good

e-ISSN: 2721-1274
p-ISSN: 2721-2718

2	18	16	20	13	33*	4	Very Good
4	49*	2	27	13	9	3	Good
5	20	31	24	20*	4	3	Good
7	11	4	18	9	58*	2	Enough
8	9	60*	16	16	0	3	Good
9	58*	16	18	7	2	3	Good
10	16	9	40*	33	2	3	Good
11	31	9	18	29*	13	3	Good
12	38	4	4	11	42*	2	Enough
13	13	42*	18	9	18	3	Good
14	4	2	56*	22	16	3	Good
15	7	42*	16	24	11	3	Good
16	7	42*	16	24	11	3	Good
17	13	13	27	11	36*	3	Good
19	24*	18	44	9	4	3	Good
20	18	67*	9	4	2	2	Enough

- Description: *Answer Key

5. Evaluation Stage

After the implementation stage of small and large class tests, the next step is the evaluation stage. The evaluation stage is the stage of the re-analyzing the results of the trial and drawing conclusion from all stages of the research. The result of the validation analysis of 20 multiple-choice test items using the Aiken's V formula have an average value 0f 0.85 and the result of the content validation test show that all test show that all test items are suitable for use in small class tests with revisions to test construction, material, and language.

In (Weisdiyanti & Juliani, 2020) test items that do not meet the material aspect are caused by (1) the compatibility between the item indicators, the cognitive level to be achieved with the question and the answer choice given, which are not appropriate, and (2) the stimulus for the item and the answer choices are not yet contextual. Test items that di not meet the construction aspect are caused by: (1) the main statements and question item formulated are not short, clear, and firm, (2) graphs and tables that are not clear and functional. Test items that do not meet the language aspect are caused by: (1) the sentences in the items used are not yet communicative, (2) there are still repetitive words in the answer choices.

In small classes, there are 3 test instruments that are invalid. Invalid test instruments must be discarded, so that from 20 questions to 17 questions will be tested in large classes. The validation of test instruments in small classes is higher than in large classes. The difference is caused by the value data and answers of test participants in small classes being more varied than those of wider field test participants. The test instrument is more valid if the test taker's scores and answers are more varied.

The validity of test instruments in small classes is higher than in large classes because the scores and answers of test takers are more varied, and the average score of test takers in small classes is higher than in large classes. In large classes, the number of questions tested was 17, with validation results of 2 invalid questions and 15 valid questions. The scores of large-class test takers were lower due to a possible cause, namely that the test takers higher level thinking abilities were lower than those of small classes.

e-ISSN: 2721-1274
p-ISSN: 2721-2718

The results of the reliability analysis on the small class test with a sample size of 10 test participants of $r11 = 0.92$ show that the HOTS based test instrument has very high reliability. The results of the reliability analysis of questions in the large class test with a sample size of 45 students amounting to $r11 = 0.808$ show that HOTS based multiple-choice questions have high reliability. All test results show that the questions are reliable in the very reliable category in the small class test and reliable in the good category in the large class test, but the test results for the small class are higher than those for the large class.

The data show that the discriminating powers of the test instrument are predominantly good. The data show that the test instruments in the large class have low discriminating power, which is more dominant than the discriminating power in the small class. The data show that the test instruments are more difficult in large classes than in small classes. A good test instrument is a test instrument with a medium level of difficulty (Widiyanto, 2018) and (Arikunto, 2008). The difficulty level of the questions in the small class test results is 95% medium and 5% difficult. Meanwhile, the difficulty level of the questions in the large class test results was 82% medium and 18% difficult, which shows that the students higher order thinking skills in the large class are still low on average.

Data analysis shows that the distractor test instrument in the large class is better than the small class, and both tests have functioned as distractors. Good distractors for HOTS based test instruments are distractors that are similar to key items and demand a high level of discriminative judgment (Scully, 2017). The small class test results showed that 1 test item had a bad distractor. Data analysis shows that students responses regarding the HOTS based test instrument have presented questions with material that students have studied (Translational Dynamics material). The questions use standard Indonesian, are communicative, do not give rise to multiple interpretations, and are easy for students to understand.

The questions are presented with attractive pictures, and the question instructions are easy for students to understand. However, there are some students who do not easily understand the questions just by reading the statements and questions. Most students cannot answer HOTS based questions easily. The time provided is according to the number of questions available. HOTS based test instruments challenge students to carry them out. Most of the criticism and suggestions from students suggest that the question form is too long, detailed, and many students still find it difficult to do it.

B. Discussion

Comparison of small class and large class test results in terms of validation, reliability, discriminating power, level of difficulty, and effectiveness of question distractors shows that the small class test results are better than the large class test results. The grades obtained by students in small classes are more varied than in large classes. Analysis of the scores achieved by students from the results of the difficulty level test shows that the higher order thinking skills of students in large classes are more homogeneous and lower than those in small classes.

The results of discriminating power tests for different test items show significant differences in the discriminating power level of test instruments in large and small classes. Test instruments of large and smaller classes have a more dominant low power difference compared to different power in smaller classes. The discriminating power difference of test items in larger classes is lower because the

e-ISSN: 2721-1274
p-ISSN: 2721-2718

difference in the value of test participants in upper and lower classes is not very variable but still sufficiently able to distinguish test participants with high abilities from students with low abilities. According to Yunita (2012) in (Weisdiyanti & Juliani, 2020), the strength varies depending on the homogeneity of the test participants, so there is a difference in results if the study is used on groups of test subjects with different characteristics.

The low level of high level skills of students is due to not being used to working on HOTS questions, according to the teacher interview questionnaire analysis. HOTS based test instruments contain activities for analyzing, evaluating, solving problems, and making decisions. On the other hand, the test results of test items in the aspects of construction validation, reliability, discriminating power, and level of difficulty show that there is an interrelated relationship.

The better the discrimination and difficulty level of the test items, the more valid and reliable they will be, and vice versa. The validation tested on each test item is in the form of construction validation. The construction of test items is good if the sentences in the main statement, main question, and answer choices are clear, concise, and do not contain ambiguous sentences. The construction of test items that are clearly concise and unambiguous is a determining factor in the quality of the different power and levels of difficulty of the test items.

CONCLUSIONS

Based on the results of research and discussion, it can be concluded that the Physics test instrument based on Higher Order Thinking Skills (HOTS) in SMA on Translational Dynamics Material for the 2023/2024 academic year was developed in the form of multiple-choice questions with 20 questions in small classes and 17 questions in large classes with analysis and problem solving skills. The conclusion of the analysis and discussion results is that the test instrument is suitable for use as a measuring tool for students higher order thinking skills and has the following characteristics:

1. Valid according to material, construction, and language experts is high with an average value of 0.85 and has obtained empirical evidence through construction validation with 85% valid items in the small class test and 88% valid items in the large class test.
2. Reliable with a value of 0.92 in the very high category ($r \geq 0.70$) for small class test results and 0.80 in the high category ($r \geq 0.70$) for small class test results.
3. The average differentiating power of test items is 0.56 with 95% of items in the good category in the small class test, and the average differentiating power is 0.42 with 56% of the good category items in the large class test.
4. The difficulty level of test items in the medium category is 95% in the small class test and 82% of medium test items in the large class test.
5. The effectiveness of the test item distractor is very good, at 70%, both on small class test results and 82% of the test item distractor on large class test results.

e-ISSN:

REFERENCES

Arikunto, S. (2008). *Dasar-dasar Evaluasi Pendidikan.* Jakarta: Bumi Aksara.

Ayuningtyas, N. & Rahaju, E. B. (2009). *Proses Penyelesaian Soal Higher Order Thinking Materi Aljabar Siswa SMP Ditinjau Berdasarkan Kemampuan Matematika Siswa.*

Dewi, R. A., Sriyono. & Ashari. (2015). Pengembangan Perangkat Pembelajaran Berbasis Problem Solving untuk Meningkatkan Keterampilan Berpikir Tingkat Tinggi pada Mata Pelajaran Fisika SMA N 3 Purworejo Kelas XI Tahun Pelajaran. *Jurnal Radiasi,* 06(1): 64-70.

Hanafi. (2017). Konsep Penelitian dan Pengembangan (RnD) Dalam Bidang Pendidikan. *Sanintifica Islamica: Jurnal Kajian Keislaman,* 4(2): 129-150.

Hartini, T. I. & Martin. (2020). Pengembangan Instrumen Soal HOTS (*High Order Thinking Skill*) pada Mata Kuliah Fisika Dasar 1. *Jurnal Pendidikan Fisika,* 8(1): 19-21 http://journal.uin-alauddin.ac.id/indeks.php/PendidikanFisika

OECD (2012) *PISA 2011: Science competencies for tomorrow world,* volume 1: *Analysis. Rosewood* (Drive: OECD)

Purbaningrum, K. A. (2017). Kemampuan berpikir tingkat tinggi siswa smp dalam pemecahan masalah matematika ditinjau dari gaya belajar. *JPPM,* 10(2): 40–49.

Saputra, H. (2016). *Pengembangan Mutu Pendidikan Era Global: Penguatan Mutu Pembelajaran dengan Penerapan HOTS (High Order Thinking Skills)*. Bandung: SMILE's Publishing.

Scully, D. (2017). Constucting Multiple-Choice Items to Measure Higher-Order Thinking. Practicial Assessment, Research & Evaluation, Vol 22, No 4: 1-13 Available online: http://pareonline.net/getvn.asp?v=22&n=4

Sudijono, A. (2008). *Pengantar Evaluasi Pendidikan*. Jakarta: PT. RajaGrafindo Persada.

Sugiyono. (2015). *Metode Penelitian Kombinasi (Mix Methods)*. Bandung: Alfabeta.

Wahyuni, D. (2018). *Peningkatan Kompetensi Guru Menuju Era Revolusi Industri 4.0*. Vol. X, No. 24/II/Puslit/Desember/2018.

Weisdiyanti, N. (2020). *Pengembangan Instrumen Tes Fisika Berbasis Higher Order Thinking Skills (HOTS) pada Materi Usaha dan Energi Tingkat SMA di Kota Medan*. Skripsi, Pendidikan Fisika, Universitas Negeri Medan, Medan.

Widiyanto, J. (2018). *Evaluasi Pembelajaran (Sesuai dengan Kurikulum 2013) Konsep, Prinsip & Prosedur*. UNIPMA Press: Madiun.