Development of Critical Thinking Skills-Based Physic Test Instruments for High School

1st Nurhikmah Weisdiyanti, 2nd Rita Juliani
Email: nurhikmahweisdiyanti@gmail.com[1]

[1,2]Physics Education Department, Postgraduate, Universitas Negeri Medan, Jl. Willem Iskandar Psr V, Medan 20221, Indonesia

Abstract. Research on the development of critical thinking skills-based physics test instruments has been carried out which aims to develop instruments that meet the aspects of validity, reliability, differentiation, difficulty level and deceptive effectiveness. The type of research used is the Borg & Gall model with the stages of analyzing the problems and needs of the critical thinking test instrument, planning, developing critical thinking-based test instruments, validating material experts, construction and language, revision of expert validation, small-scale testing, small scale instrument revision, large scale test, and large scale instrument revision. Expert validation test results obtained valid test instruments. Small scale test results obtained 93% valid items, very reliable, 87% good difference power, 93% moderate difficulty and 67% good distractors. Large scale test results obtained 87% valid, reliable, 73% good difference power, 97% moderate difficulty, and 93% good distractors so that the test instrument is feasible to measure critical thinking skills.

Keywords: Critical Thinking Skills, Physics Test Instrument, Borg & Gall.

1 Introduction

Critical thinking is one of the main skills needed and needs to be provided to students along with increasing technological advances, as well as complex problems and challenges in work and daily life that are increasing [1]. Future obstacles will be greater for students because they are the younger generation. The most effective tool for preparing students to survive in the face of difficult challenges they would later encounter in the workplace is education in this situation.

The best critical thinkers can assess an argument's persuasiveness, acquire pertinent data and come to acceptable conclusions, make wise decisions taking into account a number of factors, thoroughly investigate assumptions, assess the reliability of sources, and effectively interact with others [2]. It turns out that the critical thinking abilities of students as a generation in Indonesia are relatively low, which is out of step with the country's rate of progress [3-5]. The 2018 Program for International Student Assessment (PISA) results, which show that the competency level of the majority of Indonesian students is below level 1 and ranks 74th out of 79 participating nations, demonstrate the low level of critical thinking skills [6-7]. Although the trend for PISA in Indonesia was very increasing with the PISA 2018 population coverage in Indonesia of 85% of the entire 15-year-old students, the average PISA 2018 score even declined in three areas of competence compared to 2015 [6].

There are a number of issues that contribute to student' low critical thinking abilities. The complexity of the teaching and learning process, which calls for more time, effort, and attention

ICoSTA 2022, November 01-02, Medan, Indonesia

DOI 10.4108/eai.1-11-2022.2326211

[8-10], is one of the main contributing factors. However, this process does not prioritize developing students' skills and instead focuses solely on the subject at hand [5]. Additionally, educators still hardly ever create standardized examinations that precisely assess the unique skills that children need to possess, such critical thinking ability.

In order to gather evidence of what students have learned in relation to scientific practice and a combination of concepts, educators frequently administer routine test questions from textbooks, where the use of formulas is more prevalent. They also administer objective tests and descriptions, both of which are typically ineffective [11]. Despite the fact that the questions students are given are set at a level of critical thinking that will encourage them to use more reasoning and critical thinking to solve problems, students may still find it difficult to solve problems [12]. So that pupils become accustomed to and adept in critical thinking, it is necessary to practice answering questions that need it.

The habit of completing physics tests based on critical thinking skills has been carried out in high school at SMAN 11 Medan since 2018 by providing daily test instruments, tests and tests based on critical thinking skills along with the promotion of the use of Higher Order Thinking Skills (HOTS) based questions. The results of interviews with teachers at SMAN 11 Medan that teachers have difficulty finding examples of Physics test instruments based on critical thinking skills even though they already have a module for preparing the HOTS test instrument.

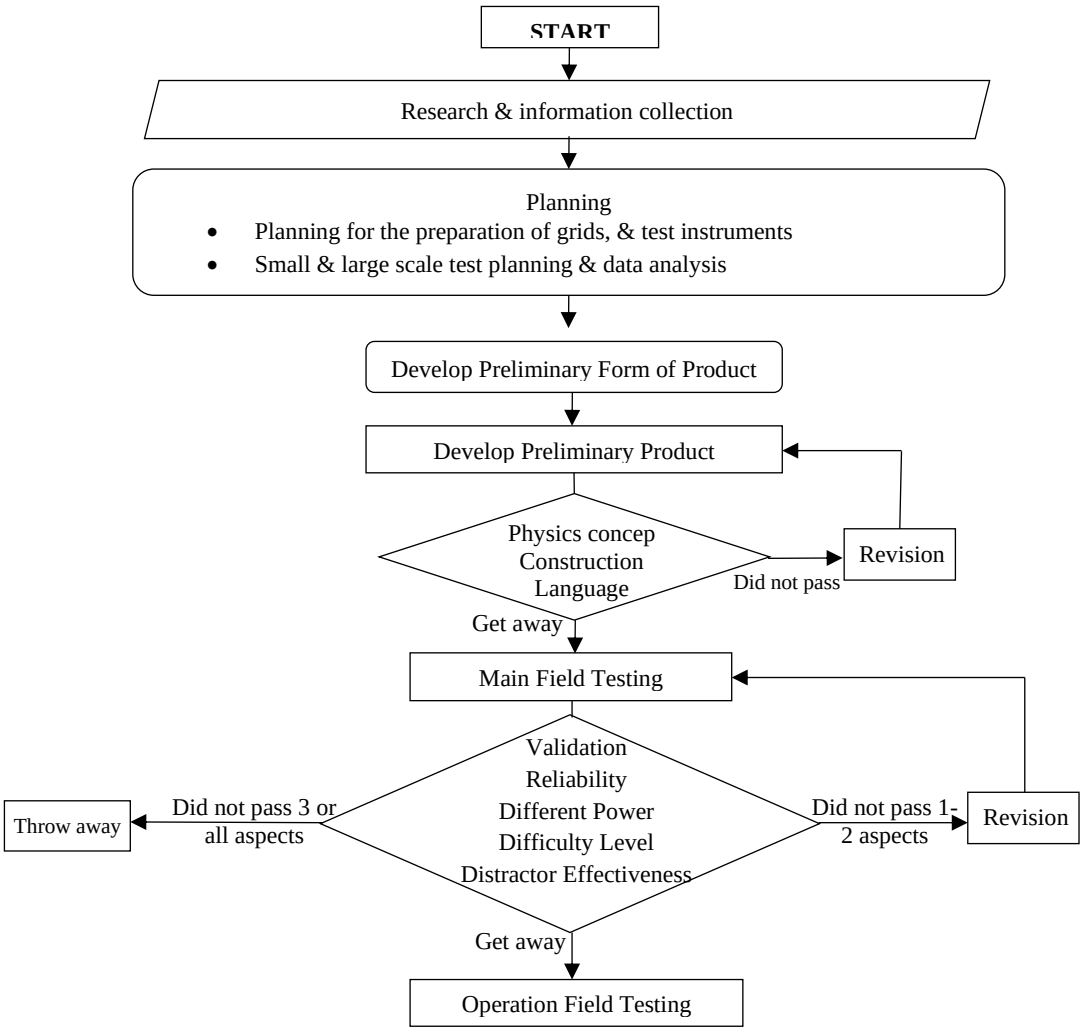
The form of the test instruments tested on practice tests, tests, midterms, and final exams at SMAN 11 Medan are multiple-choice physics test instruments, short entries and essays. The test instrument was prepared by the teacher independently. The results of the cognitive level analysis show that the test instrument presented by the teacher is at the cognitive level of C1 to C5. The test instruments tend to be at levels C1, C2, C3. The mid-semester test instrument shown by the teacher as a critical thinking skill test instrument is still at the C3 cognitive level. The test instrument is categorized as critical thinking skill if it measures the cognitive level of analytical thinking (C4-C6) of the test taker. Problems can be overcome if there are examples of appropriate critical thinking skills-based test instruments, so that teachers can apply and develop test instruments from the examples given.

The research "Development of Physics Test Instruments Based on Critical Thinking Skills for Senior High School" is important to do to improve critical thinking skills, and build students' independence to solve problems. The critical thinking skills developed in this study are from Partnership for 21st Century Skills which has identified four areas of critical thinking skills: (1) effective reasoning, (2) using systems thinking, (3) making judgments and decisions, and (4) solving problems [13].The test instrument was developed using the Borg & Gall model of research and development (R&D). The development of a physics test instrument based on critical thinking skills using the Borg & Gall model, it was found that the test instrument was suitable to be used as a tool for measuring the higher-order thinking abilities of test takers with high validity and reliability and was very effective [1]. The results of a similar study were conducted by Wakano that the test instrument developed was valid, had good reliability [14].

2 Methods

The research was conducted at SMAN 11 Medan and at SMAS Budi Satrya Medan, in the city of Medan. All participants in the study were SMAN 11 Medan test takers. The sample for this study was divided into two classes: a smaller class sample of 10 students from Class XI IPA

5 at SMAN 11 Medan and a larger class sample of 60 students from Class XI IPA 5 at SMA Budi Satrya. The study's objective test instrument based on critical thinking abilities serves as the independent variable. Validity, reliability, difficulty, discriminatory power, and distractor efficacy make up the study's dependent variables. The type of research used is the Research and Development (R&D) research model of Borg & Gall [15]. The research uses a mixed method approach, which combines qualitative & quantitative forms [16]. The research design is shown in Figure 1.



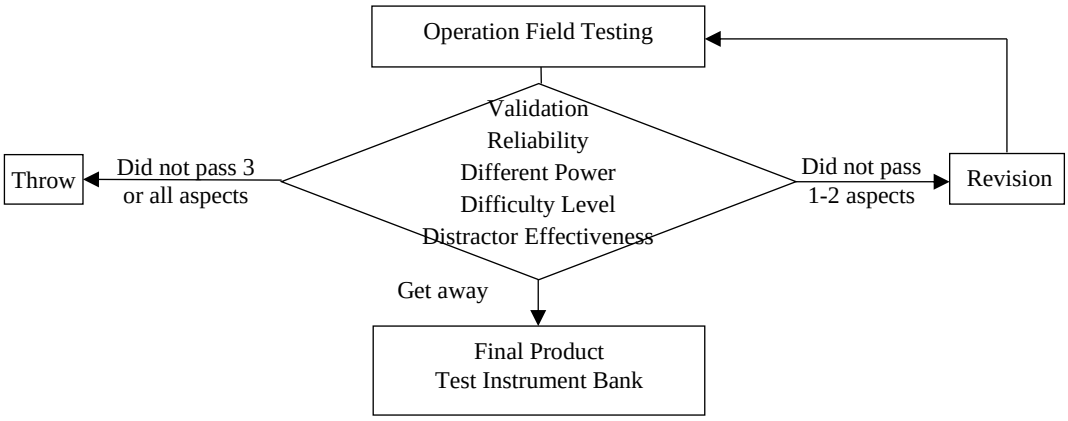


Fig. 1. Design Research and development (R&D) model Borg & Gall

The research data were analyzed qualitatively and quantitatively. Qualitative analysis was carried out using content validation determined by material expert agreement, test construction and language. Experts evaluate the test instrument using a Likert scale. The content validity index of the test items is calculated using the validity formula according to Aiken. By contrasting the item validity index of the test instrument with the value of the criteria table, the benchmark for interpreting content validity is established.

The characteristics of the test items were analyzed quantitatively on the aspects of validity, reliability, level of difficulty, discriminatory power, and distractor effectiveness of the test items described in detail as follows. Each test item is calculated for the validity of the test items by SPSS with the arithmetical validity was obtained using the product moment correlation formula [17]. If $r_{count} > r_{table}$, the test item is considered to be valid. It is necessary to revise invalid test items by improving technical proficiency in creating test instruments [18-19]. Assessment instrumentcritical thinking skills developed in the form of multiple-choice test instruments, therefore, Kuder Richardson 20 (KR-20) formula was employed to determine the instrument's reliability [20]. The benchmark for interpreting the correlation value (r_{11}) obtained by comparing the reliability coefficient with the value of the test instrument reliability criteria table [21].

Table 1. Test Item Reliability Criteria Directorate of High School Development, 2010

Reliability Coefficient	Reliability Criteria
$0{,}91 \leq r_{11} \leq 1{,}00$	Very high
$0{,}71 \leq r_{11} \leq 0{,}90$	High
$0{,}41 \leq r_{11} \leq 0{,}70$	Average
$0{,}21 \leq r_{11} \leq 0{,}40$	Low
$0{,}00 \leq r_{11} \leq 0{,}20$	Very low

The distinguishing power (DP) of test items was analyzed using SPSS [17]. Comparing the discriminatory power correlation coefficient with the value of the discriminatory index criteria table shown in table 2 serves as the standard for interpreting the discriminatory power index. Test items with positive discriminatory power should be rejected, while test items

with negative discriminatory power should either be altered or discarded. Test items with positive discriminatory power can be quite acceptable with revision [19].

Table 2. Criteria for Distinguishing Power Index

DP Coefficient	DP Criteria
$0{,}71 \leq DP \leq 1{,}00$	Very good
$0{,}41 \leq DP \leq 0{,}70$	Good
$0{,}21 \leq DP \leq 0{,}40$	Normal
$0{,}00 \leq DP \leq 0{,}20$	Not good
$0 < DP$	Bad

The test items' degree of difficulty analyzed by SPSS. A good test instrument is a test instrument with a medium category, but a test instrument that is too easy and difficult does not mean it should not be used, because it depends on the usefulness of each test instrument [17]. The higher-order thinking process of students can demonstrate understanding of information and reasoning (L3) so that the test instrument can be accepted without revision at the moderate and difficult level of difficulty. The difficulty index is classified according to the difficulty level indicator table listed in table 3 [20].

Table 3. Indicators of Difficulty Level of Test Items

Correlation coefficient	Difficulty Level Indicator
$0{,}71 \leq P \leq 1{,}00$	Hard
$0{,}31 \leq P \leq 0{,}70$	Normal
$0{,}00 \leq P \leq 0{,}30$	Easy

The distractor is considered good if the number of students who choose the distractor is the same or close to the ideal number. The effectiveness of the distractor (IP) is calculated using the equation 1. The distractor works well if more than 5% of the test takers have been selected. Distractors do not function properly if less than 5% [22].

$$IP = \frac{P}{N} \times 100\% \quad (1)$$

The criteria used to determine the quality of the test items were adapted from the Likert scale listed in table 4 [23].

Table 4. Quality Criteria for Test Items

Criteria Fulfilled	Test Item Quality	Information Revision	Enter the test Instrument Bank
4	Very good	No Revision	Yes
3	Good	Minor	Not yet
2	Fair	Major	Not yet
1	Not Good	Discard	Not
0	Bad	Discard	Not

3 Results and Discussion

The study's findings are presented as a test of critical thinking abilities on business and energy-related topics. This tool's strategic purpose is to evaluate high school students' analytical abilities using business and energy-related content. The results can be utilized as a standard for studying the phenomenon of students' abilities thanks to the test instrument's ability to represent critical

thinking abilities. The concepts or conceptual relations that students believe can be demonstrated using their analytical skills.

3.1 Description of Physics Test Instruments Based on Critical Thinking Skills

The product developed in the form of a Physics test instrument based on critical thinking skills from 10 multiple-choice test items in the question bank of SMA N 11 Medan to 15 test items in the form of multiple choice on the matter of work and energy. The test instrument is designed using a stimulus that displays concepts, visualizations, analogies, and conclusions [24] so that they can generate critical thinking skills. The test instrument is packaged into a test instrument bank with 10 critical thinking test items. The distribution of the critical thinking skills test instrument and the cognitive level of the test items that have been developed are listed in table 5.

Table 5. Cognitive Level Developed critical thinking skills test instrument

Cognitive Level	Test Instrument Number	Percentage of the number of test items
Analyze (C4)	1, 2, 5, 9, 10, 12	40%
Evaluate (C5)	3, 7, 8, 11	26.7%
Create (C6)	4, 6, 13, 14, 15	33, 3%

The developed test instrument raises the content of work and energy combined with other mechanical concepts. Each question is combined with various other materials that are still in the realm of mechanics. This combination of various materials is used to familiarize students with complex thinking. The combination of several materials used are straight motion, parabolic motion, circular motion, translational dynamics, rotational dynamics, simple harmonic motion, and dynamic fluid. The phenomenon is very close in everyday life because it uses objects that are always around the city of Medan, but there are still many students who experience misconceptions [25-26]. One of the most influential factors is the mathematical formulation that is too dominant in learning. This test instrument seeks to reveal the critical power in students' thinking, how they analyze the phenomena of work and energy in life and have mastered mathematical formulas or equations. Aspects of the test instrument based on critical thinking skills on the material of effort and energy are set as indicators for the description of the assessment test listed in table 6.

Table 6. Aspects and Indicators critical thinking skills

CT Aspect	Specific Aspect	Work & Energy Critical Thinking's Aspect Domain
Hypothesis testing	Interpret the relationship between variables	Show the relationship between variables in the work and energy
	Recognize the need for more information in drawing conclusions	Show the lack/adequacy of information in the form of mathematical or logical linkages between variables
	Identify when the principle of causality can and cannot be made	Show the cause and effect by looking at the speed and position of object on the type of energy appropriately.

Argument Analysis	Identify important parts of an argument	Show the cause and effect by looking at the speed and position of object on the type of energy appropriately.
	Criticize the validity of generalizations in experiments	Show the lack/adequacy of analysis/data from some of the induction arguments (generalizations) appropriately
Reasoning	Evaluate validity data	Show the relationship/trends between experimental data, the suitability between data and then concluded correctly.
	Detect ambiguity	Answer by indicating an error in the data and its cause.
Likelihood and uncertainty analysis	Predict the probability of an event	Show consideration of the most likely event based on the consideration of the factors affected
	Understand the need for additional information in making decisions	Show consideration of other influences that are not narrated in phenomena
Problem solving and decision making	Identify the best decision among several alternatives in solving problems	Show some alternative rational solutions
	evaluating solutions to problems and making strong decisions	Show and classify several alternative solutions with rational considerations to do

3.2 Result

After the initial product has been made, the next step is to test the validity of the content by the experts. The material expert checks the criteria for the test instrument which includes (a) the suitability of the items with the test objectives and the population of the test takers, (b) the accuracy of the information presented in the items, and (c) the clarity of words, phrases, diagrams of each item [27]. The test instrument was tested qualitatively through construction validation tests by experts on aspects of physics material, question construction and language.

The validation of the construction of test items by experts obtained an average value of 0.848 with very valid and valid criteria. The percentage of small-scale test results is shown in Figure 2.

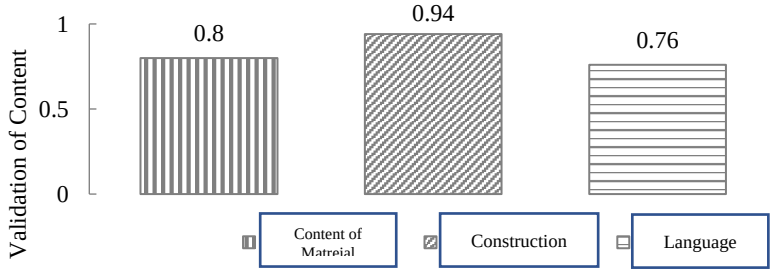


Fig. 2. Test Item Content Validation Test Results

Figure 2 shows the results of the test item validation test by experts that the test instrument is very valid in the construction aspect, valid in the content and language aspects.

All test items deserve to be tested on a small scale with 14 of the 15 test items needing revision on the material and language aspects.

The revised test instrument based on the results of the construction validation test was then carried out with a limited field test conducted on 10 test participants. The test was conducted to analyze the characteristics of the test items on the aspects of validity, reliability, discriminating power, level of difficulty and distractor effectiveness. The characteristics of the test items based on the results of the limited field test are shown in the table 7.

Table 7. Characteristics of the Test Items Based on the Results of the Limited Field Test

Test items	Validity	Reliability	Different Power	Difficulty Level	Distractor Effectiveness	Note.
1	Valid	Excellent	Good	Hard	Good	Accept
2	Valid		Good	Medium	Fair	Minor Revision
3	Valid		Good	Medium	Fair	Minor Revision
4	Valid		Excellent	Medium	Good	Accept
5	Valid		Excellent	Medium	Fair	Minor Revision
6	Valid		Excellent	Medium	Not Good	Minor Revision
7	Invalid		Bad	Medium	Good	Rejected
8	Valid		Good	Medium	Good	Accept
9	Valid		Excellent	Medium	Good	Accept
10	Valid		Excellent	Medium	Fair	Minor Revision
11	Valid		Good	Medium	Good	Accept
12	Valid		Fair	Medium	Good	Minor Revision
13	Valid		Good	Medium	God	Accept
14	Valid		Good	Hard	Excellent	Accept
15	Valid		Good	Medium	Good	Accept

The test instrument construction validation on a small scale averaged 0.693 with a V index ranging from 0.2193 to 0.9278 at a significant level of 5% with a percentage of 93% valid and 7% invalid. Table 7 shows the percentage of test item validation test results on a small scale class that 14 test items are proven valid (r_{XY} >0,4973) and 1 test item is invalid. Invalid items are then discarded.

The results of the test instrument reliability on a small scale obtained at 0.931. The data shows that the test instrument is very reliable. Instruments that have very high reliability indicate the consistency of the instrument in measuring the critical thinking skills of test takers, the level of confidence of the evaluator in placing the test instrument as the result of the evaluation and important factors in considering the results of the interpretation of the test instrument can be operationalized [22].

The average difference power of the test instrument is 0.57 with a range of different power values between 0.00 – 1.00. The data in table 7 shows that 87% of the test items are in

the very high and high category and 13% in the sufficient and not good category. Test items with poor discriminating power are invalid test items.

The data in table 7 of the 15 test items tested on the aspect of the level of difficulty shows that there are 13 test items classified as moderate and 2 test items classified as difficult and there is no test instrument that is included in the easy category. The test instrument has a range of difficulty levels between 0.2 - 0.7. The data shows that the test instrument is in accordance with the characteristics it should have, which measures students' critical thinking skills. Critical thinking skills are one of the higher-order thinking skills where the questions developed are in the category of medium and difficult level of difficulty.

The effectiveness of the test item distractors was analyzed manually using Excel. The test item distractor effectiveness test results listed in table 7 show that 66.67% of the test items have good distractors, 26.66% are sufficient and 1 test instrument is 6.67% not good. Test items with distractors that are not good, but valid will be corrected in the available answer choices section.

Overall, there was 1 test item that was discarded in the results of the limited field test data analysis. Test items received with minor and major revisions will be revised and re-analyzed through expert advice. After being valid as a whole, the test items will be analyzed for their characteristics again in a wider field test on the aspects of validity, reliability, discriminating power, level of difficulty and distractor effectiveness. The results of the wider field test are listed in table 8.

Table 8. Characteristics of the Test Items Based on the Results of the Wider Field Test

Test item	Validity	Reliability	Different Power	Difficulty Level	Distractor Effectiveness	Note.
1	Valid	Good	Excellent	Hard	Good	Accept
2	Valid		Good	Medium	Good	Accept
3	Valid		Excellent	Hard	Good	Accept
4	Valid		Good	Medium	Good	Accept
5	Invalid		Excellent	Medium	Good	Major Revision
6	Valid		Excellent	Medium	Excellent	Accept
8	Valid		Fair	Hard	Good	Minor Revision
9	Valid		Excellent	Medium	Good	Accept
10	Invalid		Not good	Hard	Excellent	Major Revision
11	Valid		Excellent	Medium	Fair	Minor Revision
12	Valid		Fair	Hard	Good	Minor Revision
13	Valid		Excellent	Medium	Good	Accept
14	Valid		Good	Medium	Good	Accept
15	Valid		Fair	Medium	Good	Minor Revision

Table 8 shows that the validation of test instrument items on a large scale averaged 0.496 which ranged from 0.084 to 0.838 with a percentage of 87% valid and 13% invalid. It is known

that 13 test items were tested valid ($r_{xy\ count} > 0{,}2108$) and 2 test items were invalid. 2 invalid test items will be discarded.

The results of the reliability test in a wider field were obtained at 0.855 with a significant level of 5%. The large-scale test data shows that after being revised, the test instrument has high reliability. The data is in line with the results of the validation test which also decreased after being revised.

The difference power of test instruments on a large scale is an average of 0.53. The data in table 8 shows that 73% of test items are in the very high and high category and 27% in the sufficient and not good category. The differentiability of the test items after being revised also decreased, according to the results of validity and reliability.

The results of the analysis of the 15 test items tested, there were 10 test items (67%) classified as moderate and 5 test items (33%) classified as difficult and no test instrument included in the easy category. The test items after being revised are more difficult than before. The increase is quite significant from 13% to 33%.

The results of the large-scale distractor effectiveness test showed that 93.33% of the test items had good distractors and 6.67% were sufficient. The quality of the distractor effectiveness of the test items increased compared to the previous 66.67% items that had good distractors.

The results of the wider field test analysis showed that 3 test items needed minor revisions, 4 test items needed major revisions and no test instruments were discarded. The test instruments that have been eligible are then entered into the bank of Physics test instruments based on critical thinking skills.

3.3 Discussion

1) Test Item Validation

The bar chart in Figure 2 shows that the test instrument is valid with revisions to the material and language aspects. The test items that do not meet the material aspects are caused by 1) The compatibility between the indicators of the test instrument, the cognitive level to be obtained with the questions, and the incorrect answer options provided are the causes of the test items that do not match the material requirements. (2) The answer options and the test's stimuli are not contextual. The sentences utilized in the test items are not yet communicative, and the terms in the answer choices are repeated, which results in the test items not meeting the language aspect.

The data in table 7 and 8 demonstrate the validity of the physics test instrument based on critical thinking abilities insofar as it can assess the critical thinking abilities of test takers from SMAN 11 Medan and SMAS Budi Satrya. There is no question about the validity of the test instrument or its accuracy in gauging students' abilities. Valid test items reflect this [28]. The construction and material of the test device cover the entire object to be measured, making it valid.

Validation of test instruments on a small scale is higher than on a large scale. The difference is because the data on scores and answers of test takers on a small scale are more varied than those of large scale test takers. The test instrument is more valid if the scores and answers of the test takers are more varied. The results of the test instrument validation are not in line with the research of [29]. The test instrument will be more valid if the number of test takers increases. The more test takers, the more varied the answers, the more valid the instrument [29].

The validity of test instruments on a small scale is higher than on a large scale because the scores and answers of test takers are more varied and the average value of small scale test takers

is higher than the large scale. The lower scores of the large-scale test takers were due to two possible causes, namely the higher-order thinking ability of the test takers was lower than that of the small-scale, and the test instruments were getting more difficult. The test instrument that has been tested on a small scale is revised so that the subject matter of the test instrument is deeper, the answer choices are more analytical and the cognitive level of the test instrument is higher so that the large-scale test instrument becomes more difficult than the small-scale instrument.

The scores and answers of the small-scale test participants are more varied than the large-scale because the construction and material of the small-scale test items cover the whole thing to be measured. The main points, questions and answer choices on a small scale are more appropriate to measure the higher-order thinking skills of test takers with a moderate level of difficulty of test items, so it is necessary to improve the construction aspect with a moderate level of difficulty of the test instrument in the revision of the large-scale test results before being included in the instrument bank test.

Reliability

Table 7 and 8 data shows that both test instruments are reliable (but the results of the small-scale test are higher than those on the large-scale. The difference is due to the value and distribution of answers of small-scale test takers more varied than large-scale test takers. The reliability of the test instrument refers to the consistency or stability of the assessment results) [30].

Instruments that have high reliability show the consistency of the instrument in measuring the higher-order thinking skills of test takers, the level of confidence of the evaluator in placing the test instrument as an evaluation result and important factors in considering the results of the interpretation of the test instrument can be operationalized [22]. A reliable instrument will get results that are not much different when used in other schools [31]. The consistency of the test instrument refers to the precision of the scores and answers of the test takers, both at SMAN 11 Medan and SMAS Budi Satrya. The data on the results of the large-scale test in the large-scale test at SMAN 11 Medan and SMAS Budi Satrya are not precise so that the test instrument that has been revised and retested on a large scale does not consistently measure the higher-order thinking skills of the test takers.

Different Power

The data in Table 7 and 8 show that the distinguishing power of the dominant test instruments is good, some still need revision and no test instruments are discarded [32]. This shows that the entire test instrument is able to distinguish test takers who have critical thinking skills from test takers who do not yet have critical thinking skills.

The results of the field test show that the test instrument on a large scale has a lower power difference which is more dominant than the difference power on a small scale. The low discriminating power of test instruments is due to test instruments that contain bias and are too difficult [33]. The results of this test are not in line with the research [32], because the more subjects, the better the differentiating power of the test items.

Difficulty Level

The data in Figure 7 and Figure 8 show that the test instruments are more difficult on a large scale than on a small scale. A good test instrument is a test instrument with a moderate level of difficulty [17]; [20]. Large-scale test instruments are more difficult than small-scale tests because the small-scale test instruments are revised on the subject of the test instrument questions with higher cognitive levels and more analytical answer choices. The main questions on the large-scale test instrument were higher because the items were revised by manipulating operational verbs to the cognitive level items at the reasoning level (L3). Distractors become more analytical than ever as distractors are revised with all of the distractors given theoretically plausible, but the key is the 'best' answer

Distractor Effectiveness

Analysis of the data in table 7 and table 8 shows that the distractor of the test instrument on the large scale is better than the small scale and both tests have functioned as a distractor with several items that need revision. This is in line with the research [34] regarding the development of a test instrument based on critical thinking skills using the R&D method. A good distractor for a test instrument based on critical thinking skills is a distractor that is similar to the key item and demands a high level of discriminatory assessment [35].

The small-scale test results obtained 1 test item that has a bad distractor. A bad distractor is caused by ambiguous sentences in the distractor [19]. Distractors whose sentences are too ambiguous need to be corrected in the language aspect so that the distractor can function properly. A distractor that can function well makes the quality of the test items good [34].

The results of small-scale and large-scale field tests show that the quality of test instruments on a small-scale field is better than large-scale based on aspects of validation, reliability, discriminating power and level of difficulty. Meanwhile, the effectiveness aspect of the test item distractor is better on a large scale than on a small scale. The quality of test items on a small scale is overall better than on a large scale. The results of the study are not in line with the research [36] regarding the development of an objective test of Physics in SMA/MA using the Borg & Gall model with the instrument results in the appropriate category in the form of aspects of validation, reliability, discriminating power, level of difficulty and distractor analysis with large-scale test results. better than the small-scale test. The quality of the test items on a large scale is less than before the small scale revision because the revised small scale test instrument is not in accordance with the aspects that need to be revised so that the construction of the test instrument becomes less harmonious and inappropriate in revising the problem.

The test item test results on aspects of construction validation, reliability, discriminating power and level of difficulty indicate an interrelated relationship. The high and low validity of the items can affect the level of reliability [37]. The better the differentiating power and the level of difficulty of the test items, the more valid and reliable the test items will be, and vice versa. The validation that is 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 statements, questions and answer choices are clear, concise and do not contain ambiguous sentences. This clear, concise and unambiguous construction of test items is a determining factor for the quality of discriminating power and the level of difficulty of the test items.

The answer choices for large-scale test items are becoming more analytical but more ambiguous than test items on a small scale, so distractors are more likely to function on a large scale. Distractors who are too deceptive make the test items less valid and reliable, more difficult and the discriminatory power is less good than before. However, the difference is not too significant, so that although the quality of the items on the large-scale test instrument decreases compared to the small-scale, the average quality of the test instrument is valid, reliable, has good discriminating power, moderate difficulty level and the effectiveness of the distractor is good so that the test instrument is feasible to measure critical thinking skills with revision of test items on aspects of validation, reliability, discriminating power and level of difficulty.

4 Conclusion

The physics test instrument based on critical thinking skills for senior high school is feasible to use with the following characteristics:

(1) Valid by experts with an average of 0.8 and has obtained empirical evidence through construction validation with 93% valid test items ($r_{count} > r_{table}$) on small-scale test results and 87% valid test items on large-scale test results.$r_{hitung} r_{tabel}$
(2) Reliable with a value of 0.931 in the very high category ($r \geq 0{,}70$) for the small-scale test results and 0.787 with the high category ($r \geq 0{,}70$) in the large-scale test results.
(3) The average differentiating power of test items is 0.57 with 87% of good category items on the small-scale test and the average difference of 0.53 with 73% of the items in the good category on the large-scale test.
(4) The level of difficulty of the test items in the medium category was 93% on the small-scale test and 67% of the test items were on the large-scale test.
(5) The effectiveness of the test item distractor was very good 67%, both on the small-scale test results and 93% on the test item distractor on the large-scale test results.

References

[1] Mappalesye, N., Sari, S.S., Arafah, K. Pengembangan Instrumen Tes Kemampuan Berpikir Kritis dalam Pembelajaran Fisika. *Jurnal Sains dan Pendidikan Fisika*. 2021; 17(1): 69-82

[2] Carlgren, L., Elmquist, M., Rauth, I. Design Thinking: Exploring Values and Effects from an Innovation Capability Perspective. *The Design Journal*. 2014; 17(3): 403–424

[3] Halpern, D. F. *Critical thinking across the curriculum: A brief edition of thought & knowledge (5th edition)*. New York: Routledge; 2014.

[4] Yamanaka, S., Inoshita, H. K., & Aehara, T. M. A Study on training of critical thinking attitude in high school chemistry through instruction focusing on argumentation. *Japan Journal of Educational Technology*, 33(4), 411– 422; 2018

[5] Negoro, R.A., Rusilowati, A., Aji, M.P., Jaafar, R. Critical Thinking in Physics: Momentum Critical Thinking Test for Pre-Service Teacher. *Jurnal Ilmiah Pendidikan Fisika Al-BiRuNi*. 2020; 9 (1): 73-86

[6] Organisation for Economic Cooperation and Development. *PISA 2018 Results Combined Executive Summaries Volume I, II & III*, Boston College, Boston; 2019.

[7] Suprayitno, T. *Pendidikan di Indonesia Belajar dari Hasil PISA 2018*. Jakarta: Pusat Penilaian Pendidikan BALITBANG KEMENDIKBUD; 2019.

[8] Greenstein, L. M. *Assessing 21st century skills: A guide to evaluating mastery and authentic learning*. California: Corwin Press; 2012.

[9] Rosefsky, A. *Learning 21st century skills requires 21st century teaching*. Phi Delta Kappan. 2012; 94(2), 8–10.

[10] Stobaugh, R. *Assessing critical thinking in middle and high schools: Meeting the Common Core*. New York: Routledge; 2013.

[11] Leverty, J.T., Caballero, M.D. Analysis of the Most Common Concep Invenories in Physics: What are We Assessing. *Phys. Rev. Phys. Educ. Res.* 2018; 14, 010123

[12] Situmorang, H.F., Bunawan, W. Pengembangan Instrumen Tes Untuk Mengukur Kemampuan Pemecahan Masalah Siswa Pada Materi Gerak Parabola. *Jurnal Inovasi Pembelajaran Fisika*. 2022; 10 (3): 44-53

[13] Partnership for 21st Century Skills. *P21 Common Core toolkit: A guide to aligning the Common Core State Standards with the Framework for 21st Century Skills*. Framework; 2011.

[14] Wakano, R., Isnaeni, W., Ahmadi, F. Developing Instruments Assessment of Students' Critical Thinking and Communication Skills in Biology Learning Using Hybrid Learning Models in 3T Areas. *Journal of Educational Research and Evaluation*. 2022; 11 (1): 93-102

[15] Borg, W. R. and Gall, M.D. *Educational research: an introduction 5th*. Longman, New York; 1989.

[16] Creswell, J. W. *Steps in conducting a scholarly mixed methods study*. University of Linkoln, Lincoln; 2013.

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

[18] Marthunis, Khaldun, I & Zulfadli. Analisis Kualitas Item tes Ujian Semester Genap Mata Pelajaran Kimia Kelas X MAN Model Banda Aceh Tahun Pelajaran 2014/2015 Menggunakan Program Proanaltes. *Jurnal Ilmiah Mahapeserta tes Pendidikan Kimia*. 2015; 1 (4) :70-78.

[19] Windarto, F., Martubi. Analisis Item tes Ujian Akhir Semester Genap Mata Diklat Dasar-Dasar Mesin (Test Item Analysis of Even Semester Final Exam in The Basic of Engine Subject). *Jurnal Pendidikan Teknik Otomotif*. 2017; 9 (2): 132-143

[20] Arikunto, S. *Dasar - Dasar Evaluasi Pendidikan*. Bumi Aksara, Jakarta; 2008.

[21] Kadir, A. Menyusun dan Menganalisis Tes Hasil Belajar. *Jurnal Al*: *Ta'dib*. 2015; 70-81

[22] Sukardi. *Evaluasi Pendidikan Prinsip dan Operasional*. Bumi Aksara, Jakarta; 2008.

[23] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. Alfabeta, Bandung; 2017.

[24] Schraw, G., Robinson, D. R. (Eds.). *Assessment of higher order thinking skills.* North Carolina: Information Age Publishing, Inc; 2011.

[25] Triyani, G., Danawan, A., Suyana, I., Kaniawati. An investigation of students misconceptions about momentum and impulse through interactive conceptual Instruction with computer simulation. *In Journal of Physics:* Conference Series. 2019; 1280 (52008)

[26] Kumar, R., Zhi-gang, Z., Livadiotis, G. The learning reconstruction of particle system and linear momentum conservation in introductory physics course the learning reconstruction of particle system and linear momentum conservation in introductory physics course. *Journal of Physics: Conference Series.* 2016; 739.

[27] Wadana, I W. *Modul Penyusunan Instrumen tes Higher Order Thinking Skills (HOTS)*. Direktorat Pembinaan SMA, Jakarta; 2017.

[28] Sudijono, A. *Pengantar Evaluasi Pendidikan*. Jakarta: PT. RajaGrafindo Persada; 2008.

[29] Afriani, E., Maria, H.T., Oktaviany, E. Pengembangan Tes Higher Order Thinking Skills Materi Gerak Lurus Berubah Beraturan untuk SMA. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa*. 2019; 8 (3) : 1-12

[30] Arifin, Z. Kriteria Instrumen dalam suatu Penelitian. *Jurnal THEOREMS (The Original Research of Mathematics)*. 2017; 2 (1) : 28-36

[31] Marwan, M., Khaeruddin, Amin, B.D. Pengembangan Instrumen Asesmen Higher Order Thinking Skills (HOTS) Pada Bidang Studi Fisika. *Prosiding Seminar Nasional Fisika PPs UNM*. 2020; 2: 116 - 119.

[32] Budiman, A., Jailani. Pengembangan Instrumen Asesmen Higher Order Thinking Skills (HOTS) Pada Mata Pelajaran Matematika SMP Kelas VIII Semester 1. *Jurnal Riset Pendidikan Matematika*. 2014; 1 (2) : 139-141

[33] Sarea, M., Sayahrul, Hadi, S. Analisis Kualitas Instrumen tes UAS Mata Pelajaran Kimia SMA di Kabupaten Gowa. *Jurnal Evaluasi Pendidikan*. 2015; 3 (3) :35-43.

[34] Sari, D.R.U., Wahyuni, S., Bachtiar, R.W. Pengembangan Instrumen Tes Multiple Choice High Order Thinking Padapembelajaran Fisika Berbasis E-Learning di SMA. *Jurnal Pembelajaran Fisika*, 2018; 7(1): 100-107

[35] Scully, D. (2017). Constructing Multiple-Choice Items to Measure Higher-Order Thinking. *Practical Assessment, Research & Evaluation*, 22 (4): 1-13

[36] Harahap, W. *Pengembangan Tes Objektif Higher Order Thinking Skill (Keterampilan Berpikir Kritis) Fisika Materi Suhu dan Kalor di SMA/MA. (Studi Kasus di MAN 2 Model Medan)*. Skripsi, Pendidikan Fisika, FMIPA, Universitas Negeri Medan; 2019.

[37] Aisah, S. Pengembangan Instrumen Penilaian Higher Order Thinking Skills (HOTS) Pada Mata Pelajaran Korespondensi Kelas X OTP di SMK Negeri 1 Jombang. *Jurnal Pendidikan Administrasi Perkantoran*, 2020; 8 (1):146-15