P-ISSN 2746-9034 || E-ISSN 2774-5066
**Indonesian Physics Education Research**
Volume 03, No. 01, June 2022, 33-43

# THE DEVELOPMENT OF HIGHER ORDER THINKING SKILLS (HOTS) TEST INSTRUMENT IN SENIOR HIGH SCHOOL

Hari Otta[1], Rita Juliani[2]
[1,2] Universitas Negeri Medan, Medan 20113, Indonesia
Email: hariottas@gmail.com

***Abstract***

The research aims to develop the objective physics instrument test based Higher Order Thinking skills (HOTS) that already meet the qualification in aspects of validity, reliability, distinguishing power, difficulty level, and effectiveness of detractors. The type of research that is used in research is Research and Development (R & D), the Borg & Gall models which consists of 7 stages, namely: 1) Research and information collecting, 2) Planning, 3) Development preliminary of products, 4) Preliminary field testing, 5) main product revision, 6) Main Field Test, and 7) Main product revision. Subject of research is student class X in SMA Swasta St. Lusia Sei Rotan who totaled 10 people on the preliminary field testing and student class X in SMA Swasta St.Lusia Sei Rotan and SMA N 2 Percut Sei Tuan who totaled 50 people on the main field testing. Results of the study showed that the validation of the contents by expect judgment is valid with the average value is 87.6 and has been getting evidence empirically through validation construction item test with the average value is 3.9 ( calculate > t-table , t-table = 2.01), the reliability of item test is 0.87 with high category ( $r$ ≥0,70 ), the distinguishing power have the average value is 0.45 with a category well, the average difficulty level is rather difficult with the range value is 0.2-0.73, and the effectiveness of distractors were function, so that instrument test decent used to measure higher order thinking skills of student.

**Keywords:** Higher Order Thinking Skills, Instruments Test, Work and Energy

Physics is the science that is the basis of science and is the basis for all discipline of science. Genetic engineering and technology also make physics the basis of its development (Young and Freedman, 2007). To meet the needs of the revolution era 4.0, based on the syllabus by curriculum 2013 of physics subjects for the senior high school issued by the minister of education and culture states that learning physics in senior high school aims to train students to master knowledge, master concepts and principles of physics, have scientific skills, have science process skills, and have critical and creative thinking skills. So, to achieve this, students of senior high school cannot only have Lower Order Thinking Skills (LOTS), but must also be able to achieve Higher Order Thinking Skills (HOTS) (Kusuma et al, 2017).

The survey results of Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA) show that the ability of Indonesian students to think scientifically is low. That is because students are not well trained in solving HOTS. Based on the PISA reported by the Organization for Economic Co-Operation and Development (OECD), Indonesia is ranked 64 out of 65 countries (OECD 2012). While for PISA in 2015, Indonesia received an average score of 403 for science (third from bottom), 397 for reading (last place), and 386 for mathematics (second from bottom) from 72 participating countries. Even though

the increase in Indonesia's achievements is quite significant compared to the results in 2012, the overall achievements are still below the OECD country average.

The results of interview at SMA St. Lusia Sei Rotan, it was found that Mr. Sumitro, a physics teacher at the school, was still unable to make HOTS-based questions properly. The results of data analysis from questionnaires that have been distributed obtained 78% of students choose that the questions used by teachers can directly apply the formulas in the book. The results of the data analysis of semester exam questions about work and energy do not have an item that truly measures the ability of students' HOTS, whereas in the grid given by the teacher there should be 33.33% (5 out of 15) Multiple choice questions about work and energy given have cognitive levels C4, C5, and C6 that can measure students' high-level thinking skills (HOTS). So, it can be concluded that in school students are only accustomed to taking tests based on LOTS which have cognitive levels of C1, C2, and C3.

The result of Research by Kusuma, et al (2017) on the development of HOTS assessment instruments in physics learning using the Borg & Gall method shows that HOTS-based instruments that are made can help students practice their Higher Order Thinking Skills. The problem that was developed was declared valid. Reliability for multiple choice and essay questions is high. Distinguishing power for multiple choice questions is accepted, for essay questions most are accepted even though some are revised and rejected. Afriani, (2019) also conducted the same research and it was found that the test instrument for senior high school was appropriate to be used as a measure of students' Higher Order Thinking Skills with high validity and reliability, moderate level of difficulty and level of text readability in accordance with the measured level. And Lindawati, et al (2016) found that the Authentic Assessment Instrument to Measure Higher Order Thinking Skills for Students based on the validation results of expert lecturers, assessment experts got grades in reasonable categories, material experts got very decent category values, and language experts from three senior high school educators got a very decent category value. Based on the problem above that the purpose of development of physics test instrument in this research to evaluate the validity, reliability, distinguishing power, difficulty level, and effectiveness distractor of test instrument towards HOTS as instrument test for learning for higher students on physics learning.

## METHODS

Subjects in this study consisted of Preliminary Field Testing, namely students of class X MIPA 1 in SMA St.Lusia Sei Rotan on 2019/2020 school year with 10 students. Main Field Test, namely students of class X MIPA 1 in SMA St.Lusia Sei Rotan and students of class X IPA 3 and X IPA 5 in SMA N 2 Percut Sei Tuan on 2019/2020 school year with 50 students.

Type of research is a research and development (R&D) research. The product developed is a physics test instrument to train students' higher order thinking skills (HOTS). The instrument development research uses the Borg & Gall research model which consists of seven development steps, namely: Research and Information collection; Planning; Develop Preliminary form of Product; Preliminary Field Testing; Main Product Revision; Main Field Testing; and Operational Product Revision. For details can see in the picture 1.

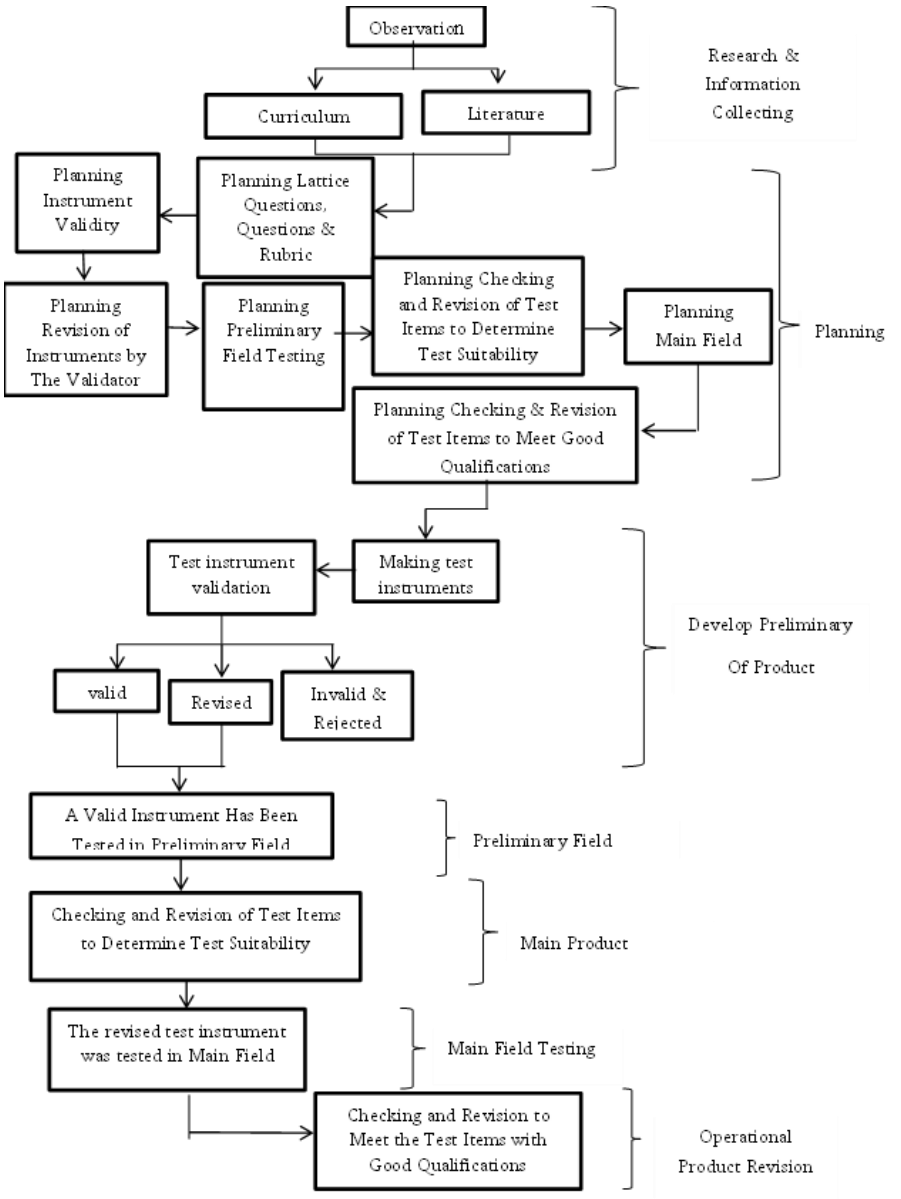


Figure 1. Research Flow

## RESULT AND DISCUSSION

### *Result*

Type of research researcher has done is Research and Development (R&D) uses the Borg & Gall research model. Research and information collecting step is done with some ways, they are literature study, interview, questionnaire data analysis, and analysis of used tests. In literature study activity, research collects some sources about the research that is done, for instance articles, books, thesis, and scientific journals. From literature study activity, it is known, that HOTS test is not enough in field. After getting the information, researcher visited one school in Deli Serdang, that is SMA Swasta St. Lusia, Sei Rotan, to have an interview with a teacher, that is in physic. This continues with last examination work and energy, that is used in that school. Analysis result of the questionnaire, researcher got 78% of students choose, that the tests, which are made by the teacher can be applicated to the formulas that are in the book.

Research and Development is research, that is done to develop and to test the effectiveness of product. Product of this research is Higher Order Thinking Skills (HOTS) test instrument based of work and energy, that is 30 items multiple choice. There are 2 cognitive processes in these tests, they are analyzing (C4), and evaluating (C5). Grid of test C4 amounts of 17 tests, and C5 amounts of 13 tests. Research instrument that is used is validation test sheet by specialist, a set of HOTS test manuscript, that contents multiple choice tests with 5 alternative answers (a, b, c, d, e), student answer sheets, and key test of ability of Higher Order Thinking Skills of work and energy in Senior High School.

Then, test instrument is given to experts to be validated using validation sheet, that has been made in the previous step. The Experts are 2 lecturers in Physics department in State University of Medan and 1 as a teacher of Physics in SMA Swasta St. Lusia, Sei Rotan. Test validation by specialist including material aspect, construction, and language. Based on expert validation using the validity index of Linkert scale items are known that 24 items are very valid (very good to use) and 6 items is valid (may be used after minor revisions) with an average validity of the test items of the instrument is 87.6.

After being given input by experts, the test instrument was revised and then a preliminary field test was conducted at SMA Swasta St. Lusia Sei Rotan, which is class X with sample are 10 students. Test used in the preliminary field testing are the 30 questions. And Main Field Test was conducted at class X MIPA 1 in SMA St.Lusia Sei Rotan and at class X IPA 3 and X IPA 5 in SMA

N 2 Percut Sei Tuan on 2019/2020 school year with 50 students. Test instrument in the main field testing are the 29 questions. Test instrument qualifications are described as follows:

1. Validation

Out of 30 items in preliminary field testing, 24 (80%) items were valid and 6 (20%) items were invalid. Out of 29 items in main field testing 27(93%) items were valid and 2 (7%) items were invalid. The result of validity can be seen in the table 1:

Table 1 Validity of Test instrument

| Item Status | Preliminary Field Test | | | Main Field Testing | | |
|---|---|---|---|---|---|---|
| | Item Question | Total | Percentage (%) | Item Question | Total | Percentage (%) |
| Valid | 1, 2, 3, 5, 7, 8, 9, 10, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28 | 24 | 80 | 1, 2, 3, 4, 5,6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28,29 | 27 | 93 |
| Invalid | 4, 6, 11, 12, 29, 30 | 6 | 20 | 7, 20 | 2 | 7 |

2. Reliability

Data show's reliability of preliminary field test is 0.92 (very high) and reliability of main field testing is 0.87 (high). The test items can be said to be reliable if they have a reliability greater than 0.70.

3. Distinguishing Power

A good test instrument requires distinguishing power to differentiate intelligence between students. Data is processed manually using Microsoft Excel. The results of the distinguishing power tests of shown in table 2.

Table 2 Distinguishing Power Test Instrument

| Distinguishing Power Coefficient | Criteria | Item questions of Preliminary Field Testing | Item questions of Main Field Testing | Decision |
|---|---|---|---|---|
| 0,71 ≤ DP ≤ 1,00 | Excellent | 1, 3, 15, 16, 18, 24, 26 | 9,11,12,15,23,26 | Accepted |
| 0,41 ≤ DP ≤ 0,70 | Good | 2, 5, 7, 8, 9, 10, 13, 14, 17, 19, 20, 22, 23, 25, 27, 28 | 1,2,3,4,5,6,8,10,13,14 | Accepted |
| 0,21 ≤ DP ≤ 0,40 | Enough | 11, 21 | 16,17,18,19,22,24,25 | Minor Revised |
| 0,00 ≤ DP ≤ 0,20 | Bad | 4, 6, 29, 30 | 27,28 | Major Revised |

| | | | | |
|---|---|---|---|---|
| 0 < DP | Very Bad | 12 | 7,21,29 | Rejected |

4. Difficulty Level

Difficulty level analysis of the questions is needed to find out the questions in the category of easy, medium, and difficult. Data is processed manually using Microsoft Excel. Percentage of difficulty test level of the preliminary field test can be seen in Table 3.

Table 3 Difficulty Level of Preliminary Field Test

| Correlation coefficient | Criteria | Item questions of Preliminary Field Testing | Item questions of Main Field Testing | Decision |
|---|---|---|---|---|
| 0,71≤P≤1,00 | Easy | 6, | 20 | Rejected |
| 0,31≤P≤0,70 | Rather Difficult | 1,2,3,5,7,8,9,10, 14, 15, 17, 18, 19, 20, 21, 22, 26, 27, 30 | 1,2,3, 4,5,6,9,10, 11,13,14,15. | Accepted |
| 0,00≤P≤0,30 | Difficult | 4, 11. 12, 13, 16, 23,24, 25, 28, 29 | 16,17,19,23,24,25, 26,27,28,29 | Accepted |

5. Effectiveness Distractor

Analysis of the effectiveness of the distractor is done by taking data directly from google form analysis. Question items are considered good if you have an effective contractor outwitting students. Distracters function well if more than 5% of followers have been chosen (Sudijono, 2012). In preliminary field testing that 26 of the 30 items have met the effectiveness of a good distractor so that it does not need to be revised, while the other 4 items do not meet the effectiveness of a good distractor so that revisions are needed, namely at numbers 4, 6, 2, and 30 in order to proceed to the next stage. more extensive field testing. In Main Field Testing 26 of the 29 items have met the effectiveness of a good distractor so that it does not need to be revised, while the other 3 items do not meet the effectiveness of a good distractor so that revisions are needed, namely at numbers 2, 20, and 29.

## *Discussion*

1. Analysis of Validity

Test items are said to be valid if t-calculate > t-table. The results of the validation analysis of the questions in the preliminary field testing showed the valid questions were 80% and invalid questions 20%. The results of the validation analysis of the questions in the main field testing showed the valid questions were 93% and invalid questions 7%.

The results of the study are in accordance with the theory of validity according to Sudijono (2012) that the items that have high validity reflect the questions already have reliability and there

is no doubt the accuracy of the questions in measuring the ability of students. This is similar to the opinion of Gronlund (2009) which states that validity is the accuracy of interpretation obtained from the assessment results. The problem is valid because the construction and material really cover the whole that you want to measure. Items with low or invalid validity need to be revised by increasing technical mastery of ways of making test questions (Windarto & Martubi, 2017).

2. Analysis of Reliability

The results of the reliability analysis of the questions in the preliminary field testing to a value of 0.92 indicating that multiple choice questions based on HOTS have high reliability. The results of the reliability analysis of the questions in the main field test with a value of 0.87 indicate that the multiple-choice questions based on high reliability HOTS. The second data of the reliability test which was processed using the Kuder Richardson 20 (KR-20) formula for preliminary field testing from 30 items and main field test from 29 items shows that HOTS-based physics questions in multiple choice form can measure the students' answers in different situations.

This is similar to the opinion of Reynold (2006) in Arifin (2017) which states that reliability refers to the consistency or stability of the results of the assessment. Instruments that have high reliability mean that the results of instrument interpretation show better consistency (Sukardi, 2008). Also believes that a reliable instrument implies that the instrument used is sufficiently robust to be used in retrieving research data, so as to be able to uncover reliable data on the results.

3. Analysis of Distinguishing Power

The results of the analysis of distinguishing power of item test in preliminary field testing showed 77% item accepted, 20% item revised, and 3% rejected. The results of the analysis of the distinguishing power of item test in the main field test showed 87% item accepted, 10% item revised, and 3% item rejected. Questions that have been accepted for distinguishing power indicate that the test instrument is able to distinguish test participants who have critical thinking skills from participants who do not have critical thinking skills.

The distinguishing power depends on the homogeneity of the test takers. This limitation causes differences in the results if the study is used in a group of test participants with different characteristics of the experimental subjects. A value of D = 1 means that the entire upper group can answer the questions correctly, while the whole lower group answers incorrectly. If students in the upper and lower groups both answer true or false (D = 0), it means that the problem has no difference. A question that has a negative value (D = -1), meaning that all lower grade students

answer right and all upper-class students answer wrong. Problems that have a negative D value should be discarded (Arikunto in widiyanto, 2018).

4. Analysis of Difficulty Level

The result of difficulty level of items in preliminary field testing there are 3% of items classified as easy, 67% items about categorized as medium, and 30% of items classified as difficult. The results of the difficulty level of item test in main field test were 3% of items classified as easy, 74% included in the Rather Difficult category, and 24% were in the difficult.

The difficulty levels range from 0.00 to 1.00. The greater of difficulty levels obtained from the calculation results, it means that the problem is easier and the item test must be revised (Arifin, 2017). Widiyanto (2018) argues that although good questions are medium category questions, questions that are too easy or too difficult do not mean they should not be used, because it depends on the use of each question. Learners are asked to transfer information from one context to another to process and apply information, see the linkages between different information, use information to solve problems, critically study / examine ideas or ideas and information (Ministry of Education and Culture, 2017), so researchers making decisions for questions that are accepted without revision is a matter of rather difficult and difficult.

5. Analysis of Effectiveness Distractor

The results of effectiveness of distractor in preliminary field testing was 87% accepted and 13% revised, while the main field test was 90% accepted and 10% revised, it shows that the effectiveness of the distractor is functioning well. Deception is said to function if selected by at least 5% of the number of students who take the test and chosen by students who lack mastery of the exam material (Ramadhan et al, 2019). Deception is more chosen by students in the smart category so it can be said that the deception is misleading (Windarto and Martubi, 2017). The results of the test and analysis of the effectiveness of distractor showed that there were no items that had misleading deception on the physics-based test instrument.

6. Analysis of Item selection

The results of the analysis of the quality of test instrument can be determined from the results of the analysis of validation, reliability, different power, level of difficulty and effectiveness of the contractor. Results of preliminary field testing, namely: (a) 23 questions were accepted without revision by category valid, different power is good and excellent, reliability is very high, rather difficulty level from rather difficult to difficult and effectiveness of distractor is very good, (b) 6 questions accepted with revision, (c) 1 questions were rejected that is number 6 with an invalid

category, the distinguishing power was bad, Difficulty level was easy and the effectiveness of the distractor was bad. Then the instrument test was revised and continued to the main field testing.

The results of the main field test, namely: (a) 24 questions were accepted without revision by category valid, different power is good and excellent, reliability is very high, rather difficulty level from rather difficult to difficult and effectiveness of distractor is very good, (b) 4 question accepted by revision (c) 1 questions were rejected that is number 20 with an invalid category, the distinguishing power was bad, the difficulty level was easy and the effectiveness of the distractor was enough. The finally researcher have 28 item tests can measure HOTS of students.

The results of the comparison of the two tests that the results of the main field test are better than the results of the preliminary field testing, according to the research of Harahap (2019) concerning the development of objective tests of Higher Order Thinking Skills (HOTS) Physics in senior high school using the Borg & Gall with the results of instruments in the proper category in the form of aspects of validation, reliability, different power, difficulty level and deception analysis with broader class test results is better than limited class test. Sanjaya (2013) in Afriani, et al (2019) states that the more subjects the better the product produced.

## CONCLUSION AND SUGGESTION

Conclusions of the results of the analysis and discussion of the development of test instruments that the physics test instrument based on Higher Order Thinking Skills (HOTS) at the Senior High School has met the requirements and is suitable for use as a measurement tool for higher order thinking skills with characteristics:

1. Validation by expert judgment is very valid with an average value of 87,6 and has obtained empirical evidence through the validation of item construction with 80% valid items in the preliminary field test and 90% valid items in the main field test.
2. The reliability of the HOTS-based test instrument for work and energy at the SMA level is 0.92 with a very high category ($r \geq 0.70$) in a preliminary field test and 0.87 with a high category ($r \geq 0.70$) in the main field test
3. The distinguishing power of the HOTS-based test instrument for work and energy at the SMA level has an average 0.6 with 87% of the categories accepted in the preliminary field test and the average 0.5 with 77% of the categories accepted in the main field test.
4. The difficulty level of the HOTS-based test instrument for work and energy at the SMA level has 30% in the difficult category and 67% in the rather difficult category in the preliminary

field test and 21% in the difficult category and 76% in the rather difficult category in the main field test.

5. Effectiveness of distractor the HOTS-based test instrument for work and energy at the SMA level is 87% accepted in the preliminary field test and 90% in the main field test.

## ACKNOWLEDGMENTS

I would like to express my appreciation for Dr. Rita Juliani, M.Si. as my supervisor who has provided guidance and advice to author so this research can be carried out. Also, in addition I'd like to thank SMA S St.Lusia sei Rotan and SMA N 2 Percut Sei Tuan who had been willing in collaboration with author as subject in this my research.

## REFERENCES

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

Arifin (2017). Prosedur Penelitian Suatu Pendekatan Praktik. Jakarta: Rineka Cipta.

Kusuma, M.D., Abdurrahman, U.M., & lSuyatna, A. (2017). The Development of Higher Order Thinking Skills (HOTS) Instrumen Assessment in Physics Study. IOSR Journal of Research & Method in Education, 7 (1): 26-32

Lindawati, L., Saregar, A., & Yuberti, Y. (2016). Pengembangan instrumen authentic assessment untuk mengukur higher order thinking skills peserta didik. In Seminar Nasional Pendidikan (pp. 140-149).

OECD. (2012). PISA 2015 Result in Focus International Resultsin Science. Boston: The TIMSS & PIRLS International Study Center.

Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure the higher order thinking skill in physics. European Journal of Educational Research, 8(3), 743-751.

Reynold (2006). Analisis kuantitatif Instrumen Penilaian (Panduan peneliti, Mahasiswa, dan Psikometrian). Yogyakarta: Prama Publishing.

Sanjaya (2013). Dasar-dasar Evaluasi Pembelajaran. Yogyakarta: Graha Ilmu

Sudijono. (2012). Metode Penelitian dan Pengembangan Research and Development. Bandung: Alfabeta.

Windarto, & Martubi. (2017) Modul Penyusunan Soal Higher Order Thinking Skill (HOTS), Jakarta: Direktorat Jendral Pendidikan Dasar dan Menengah Departemen Pendidikan dan Kebudayaan.

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

Young, H.D., & Freedman, R.A. (2007). University Physics 12th edition. New York: Pearson-Addison Wesl