JPF | Volume 11 | Number 3 | 2023 | 282 - 296

p - ISSN: 2302-8939
e - ISSN: 2527-4015

Jurnal Pendidikan Fisika

https://journal.unismuh.ac.id/index.php/jpf

DOI: 10.26618/jpf.v11i3.11829

Development of the HOTS Test to Measure Students' Critical Thinking Skills on Optical Instrument Materials

Hakiki Ernawati[1)*], La Maronta Galib[2)], Muhammad Anas[3)]

[1,2,3)]*Department of Physics Education, Halu Oleo University, Kendari, 93561, Indonesia*

**Corresponding author: hakikiernawati9@gmail.com*

Received: May 31, 2023; Accepted: July 26, 2023; Published: August 24, 2023

Abstract *– The objective of the study was to: 1) develop and produce a HOTS test that is valid and reliable; 2) examine the quality of the items; and 3) measure the critical thinking skills of high school students on the optical instruments material based on the HOTS test developed. Employing Research and Development (R & D) method, this study combined two approach models, namely the Wilson Model and the modified Oriondo and Antonio Model (2014) and the Successive Approximation Model (SAM) by Michael Allen and Richard Sites (2019). This study involved 37 students selected by purposive sampling as test subjects. The results of the study show that: 1) HOTS tests that were valid and reliable, where there were 31 valid questions out of the 60 items tested with a product-moment correlation coefficient of 0.72/high and a test reliability coefficient of 0.79/high; 2) the level of difficulty index of the items is in the range of 0.03-0.92, where 12 items with low discriminating power that need to be revised, while 16 items with very low discriminating power need to be replaced and the effectiveness of the distractor show that there were 13 items with options that need to be replaced; and 3) trials conducted on the HOTS test developed revealed students' critical thinking skills level, namely 16% low, 65% moderate, and 19% high. Therefore, it can be concluded that the use of the HOTS test is appropriate to measure students' critical thinking skills.*

Keywords: *critical thinking skills; higher-order thinking skills test; optical instruments*

I. INTRODUCTION

Critical thinking skills are needed by students to connect concepts and material to understand and solve problems and learn effectively in class (Nurilma et al., 2023; Sukmagati et al., 2020). Critical Thinking Skills (CTS) are a person's ability to carry out precise and directed thinking processes to draw conclusions, identify correlations, analyze possibilities, make predictions and logical decisions, and do complex problem solving (Damayanti & Kuswanto, 2020; Abidin et al., 2019; Marnah et al., 2021). Unfortunately, in reality, students' critical thinking skills are still low and one of the reasons is that the learning outcome evaluation instruments used in schools are generally still based on Lower-Order Thinking Skills (LOTS) questions and they are not used to solving Higher-Order

Thinking Skills (HOTS) questions in a learning process (Lespita et al., 2023; Shaheen, 2016).

The 2018 National Program for International Student Assessment (PISA) Indonesia report shows unsatisfactory results for Indonesian students, especially in the field of science (Setiyoningtyas & Kasmui, 2020). This is shown by the average score of Indonesian students in the 2018 PISA results in the science field of 396 and is in the 70th position out of 78 Organization for Economic Co-operation and Development (OECD) participating countries. This score is lower when compared to the 2015 PISA results, namely with an average score of 403 (OECD, 2019). This condition suggest the idea that Indonesian students need to be well prepared in dealing with PISA questions because these questions are always oriented towards solving problems, not just memorizing but solving them requires high-order thinking skills (Khaeruddin et al., 2023).

Relevant to the PISA and HOTS questions, students in schools are also required to master 21st-century skills so that each school has its role and responsibility in preparing students to face these conditions (Redhana, 2019). Unfortunately, in general, according to empirical data in the field, the average number of HOTS-based school exam questions used in Southeast Sulawesi province is 16.6% and this condition has not been made a priority to be implemented as a habit or routine in the learning process. This is shown by the questions for assessing learning outcomes which are still LOTS based, there are no new model questions, or there is no question bank related to material that is rarely discussed further in class (Nisa & Wasis, 2018; Sidik et al., 2021; Widjanarko, 2022). Many teachers still make test instruments that only measure the lower-order thinking skills of students, namely the Cognitive Process (CP) dimensions of CP-1 (remember) and CP-2 (understand), meanwhile, the questions that train higher-order thinking skills of students tend there isn't any (Litna et al., 2021).

Istiyono et al. (2014), revealed that in learning, higher-order thinking test instruments are very necessary and have conducted research regarding this matter in Class XI at High School in Yogyakarta for Physics Material, namely Motion, Force, Work, and Energy as well as Momentum and Impulse. Based on that, it is necessary to: 1) implement the HOTS test in high school; 2) hold HOTS test preparation training for educators; and 3) do further research was carried out regarding the development and analysis of HOTS question items. Likewise, Saputro & Supahar (2018) have researched the development of higher-order thinking test instruments to measure the achievement of students' physics learning outcomes in high school in Optical Material revealed that it was necessary to create a question package with anchor items to avoid collaboration between students when working on test instruments

and it is necessary to carry out further research to make a test instrument that is similar to different Physics Materials.

Research on the development of HOTS test instruments has been carried out, but the subject matter is still limited so more comprehensive empirical information about the HOTS test is needed to further complement previous studies (Saputro & Supahar, 2018). In addition, this research is intended to measure students' critical thinking skills which are integrated into HOTS questions and have not been widely used. The research model used displays innovations in the latest research procedures consisting of parts developed to make the process of developing tests and assembling questions better. Therefore, research related to the development of the HOTS test is very necessary because it aims to provide teachers with references related to similar matters so that they can familiarize students with the learning process by working on HOTS questions.

Specifically, the problem formulation in this research is about: 1) how to develop and produce a conceptually and theoretically valid and reliable HOTS test; 2) how to measure the quality of the questions in the form of level of difficulty, distinguishing power and effectiveness of distractors; and 3) how to measure students' critical thinking abilities both as a group and on each CTS indicator itself. Based on this formulation, this research aims to: 1) develop and produce a valid and reliable HOTS test; 2) check the quality of the questions; and 3) measuring high school students' critical thinking abilities on optical instruments using the developed HOTS test.

II. METHOD

This study used the Research and Development (R & D) method by combining two approach models, namely the Wilson Model and the modified Oriondo and Antonio Model (Istiyono et al., 2014) and the Successive Approximation Model (SAM) by Michael Allen and Richard Sites (Ali et al., 2021). This study involved 37 students selected by purposive sampling as test subjects. The trials was carried out at SMA Negeri 1 Kendari which is located at Jl. Mayjen Sutoyo, No. 102, Tipulu, Kec. Kendari, Kendari City, Southeast Sulawesi, 93122 which was carried out on December 7, 2022.

There were three Stages in this study, namely: 1) the preparation stage; 2) the iterative design stage; and 3) the iterative development stage. The preparation stage consisted of (a) determining the purpose of the test, (b) selecting the competencies to be tested, and (c) designing of test grids. The iterative design stage consists of (a) compiling the items, (b) making alpha product/prototype 1, (c) conducting test validity, and (d) revising the items and collecting the test. The iterative development stage consists of (a) making prototype 2/beta product, (b) implementing the trials, (c)

analyzing the test characteristic, and (d) revising the items and collecting the test to produce a standard test/gold product accompanied by dissemination which can be seen in Figure 1.

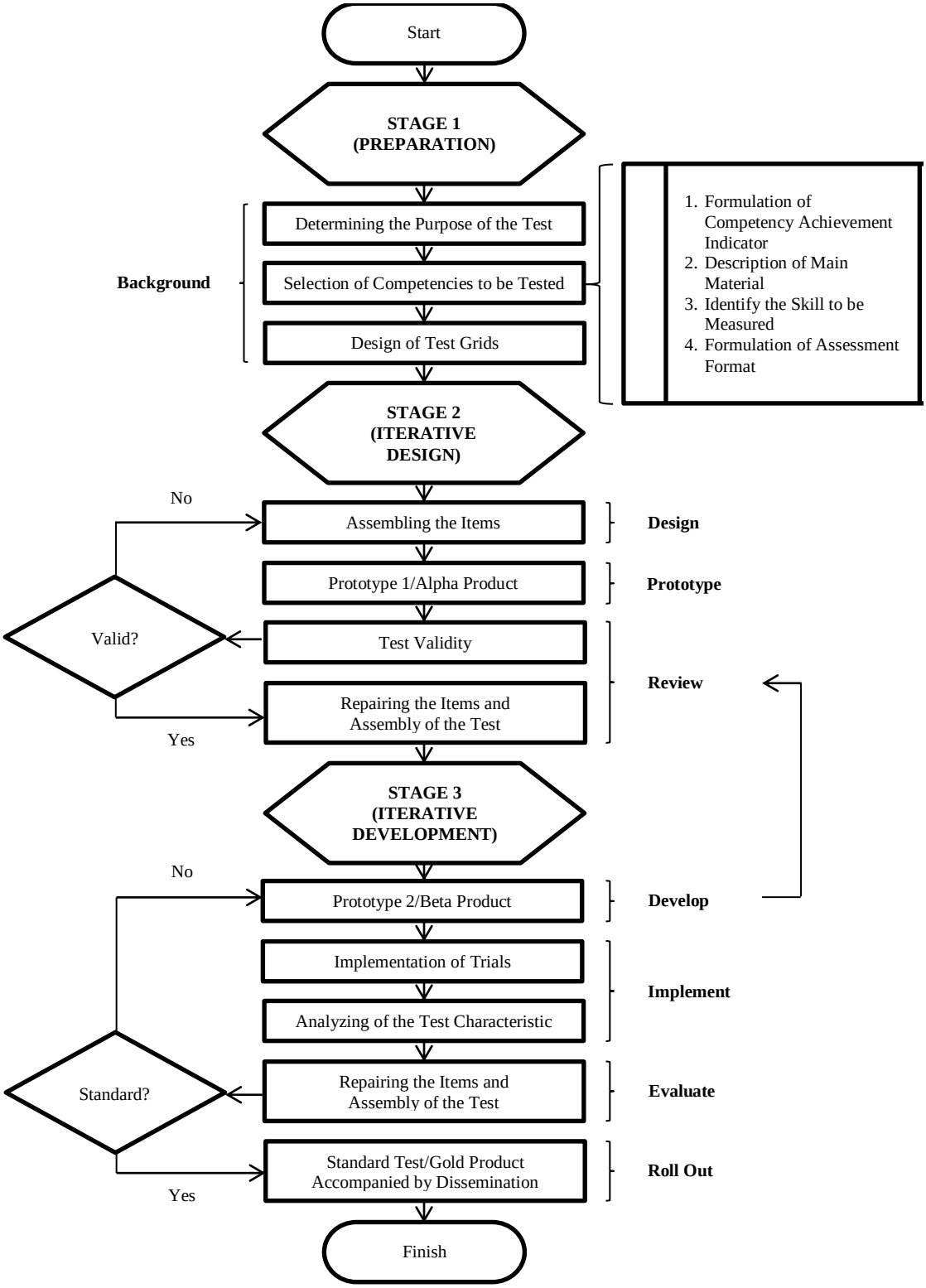


Figure 1. Research flowchart

This study used expert validation sheets that assess the material, language, and construct aspects as a test of content and construct validity by experts which were then tested empirically, namely testing the validity, reliability, and quality of the items in the form of tests of difficulty level, discriminating power, and the functioning of the distractor and the analysis of the raw data using the help of Anates V4 software.

1. Validity Test

According to Arikunto (2012), data analysis from the expert validation sheet used Aiken's *V* index calculation with the criteria and category as shown in Table 1.

Table 1. Aiken's *V* criteria and category

No	Aiken's *V* Criteria	Category
1	$V < 0.4$	Low
2	$0.4 \leq V \leq 0.8$	Moderate
3	$V > 0.8$	High

Arikunto (2012) also suggests the calculation of the validity coefficient of the items used the biserial point correlation formula with the criteria and category as shown in Table 2.

Table 2. Criteria and category of *rpbis*

No	Criteria of *rpbis*	Category
1	*rpbis count* > *rpbis table*	Valid
2	*rpbis count* < *rpbis table*	Invalid

Based on Arikunto (2012), the calculation of the validity coefficient of the test used the product moment correlation formula and the criteria and category as shown in Table 3.

Table 3. Coefficient interval and category of r_{xy}

No	Coefficient Interval of r_{xy}	Category
1	0.81 – 1.00	Very High
2	0.61 – 0.80	High
3	0.41 – 0.60	Moderate
4	0.21 – 0.40	Low
5	0.00 – 0.20	Very Low

2. Reliability Test

Based on Arikunto (2012), the calculation of the reliability coefficient of the test used Kuder-Richardson's formula with the criteria and category as shown in Table 4.

Table 4. Coefficient Interval and Category of *KR-20*

No	Coefficient interval of *KR-20*	Category
1	0.00 – 0.19	Very Low
2	0.20 – 0.39	Low
3	0.40 – 0.59	Moderate
4	0.60 – 0.79	High
5	0.80 – 1.00	Very High

3. Difficulty Level Test

Arifin (2012) claims that the calculation of the item difficulty level index (*p*) and the test difficulty level index (*TK*) were calculated using Equation 1. Meanwhile, the criteria and category are presented in Table 5.

$$p = \frac{\sum B}{n}, \quad TK = \frac{(WL+WH)}{(nL+nH)100} \qquad (1)$$

When,

p = item difficulty level index
ΣB = number of subjects who answered correctly
n = number of subjects
TK = test difficulty index
WL = erroneous number of subjects from the lower group
WH = erroneous number of subjects from the upper group
nL = number of subjects in the lower group
nH = number of subjects in the upper group

Table 5. Criteria and category of *p* and *TK*

No	Criteria of *p* and *TK*	Category
1	$p, TK < 0.30$	Difficult
2	$0.30 \leq p, TK \leq 0.70$	Medium
3	$p, TK > 0.70$	Easy

4. Discriminating Power Test

Based on Arifin (2012), the calculation of the item discriminating power index (*DP*) used Equation 2. The criteria and category are as presented in Table 6.

Table 6. Criteria and category of *DP*

No	Criteria of *DP*	Category
1	$DP \leq 0.00$	Very Low
2	$0.00 < DP \leq 0.20$	Low
3	$0.20 < DP \leq 0.40$	Moderate
4	$0.40 < DP \leq 0.70$	High
5	$0.70 < DP \leq 1.00$	Very High

$$DP = \frac{B_A}{N_A} - \frac{B_B}{N_B} \quad (2)$$

When,

DP = item discriminating power index
B_A = number of subjects in the upper group who answered correctly
B_B = number of subjects in lower group who answered correctly
N_A = number of upper group subjects
N_B = number of lower group subjects

5. Test the Effectiveness of Detractors

Based on Arifin (2012), the calculation of the item distractor index (*IP*) is as in Equation 3. The criteria and category are as presented in Table 7.

$$IP = \frac{P}{(N-B)/(n-1)} \times 100\% \quad (3)$$

When,

IP = item detractor index
P = number of subjects chose a distractor
N = number of subjects
B = number of subjects answered correctly on each question item
n = number of answer options

Table 7. Criteria and category of *IP*

No	Criteria of *IP*	Category
1	*IP* = 76% - 125%	Very High
2	*IP* = 51% - 75% or 126% - 150%	High
3	*IP* = 26% - 50% or 151% - 175%	Moderate
4	*IP* = 0% - 25% or 176% - 200%	Low
5	*IP* > 200%	Very Low

6. Critical Thinking Skills Test

Based on Arikunto (2012), the calculation of the CTS index of test participants as subjects is in Equation 4. The criteria and category are presented in Table 8.

$$\bar{x} = \frac{\sum x}{n}, \quad SD = \sqrt{\frac{\sum(\bar{x}-x)^2}{n-1}} \quad (4)$$

Table 8. Criteria and Category of 'Subject'

No	Criteria of x	Category
1	$x > \bar{x} + 1SD$	High
2	$\bar{x} - 1SD \leq x \leq \bar{x} + 1SD$	Moderate
3	$x < \bar{x} - 1SD$	Low

Based on Sudjana (2005), the calculation of the CTS index of test participants when they are in a certain population used Equation 4. The criteria and category are as Table 9.

Table 9. Criteria and category of 'population'

No	Criteria of μ	Category
1	$\mu > \bar{x} + \frac{z_{\alpha/2}SD}{\sqrt{n}}$	High
2	$\bar{x} - \frac{z_{\alpha/2}SD}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{z_{\alpha/2}SD}{\sqrt{n}}$	Moderate
3	$\mu < \bar{x} - \frac{z_{\alpha/2}SD}{\sqrt{n}}$	Low

When,

x = score obtained
n = number of subjects
SD = standard deviation
$\bar{x}$ = test subject's average score
μ = population mean score estimation
$z_{\alpha/2}$ = z value which gives a probability of 1-α for the two-party test

III. RESULTS AND DISCUSSION

1. Validity

The test was declared valid and with revisions for items with low validity which were then tested on test takers indicated by the Aiken's V index value of 0.69 in the moderate category and 29 items were invalid with correlation coefficients of items that were in the range -1 to +1 and a test validity coefficient of 0.72 or categorized as high. Follow-up on the results of the analysis of the validity of the items, namely valid items can be reused in the next test or stored in the question bank and invalid items can be used, but it is necessary to review them qualitatively in the form of being revised and adjusted to the construct, use in presentation of material and assessment formats that were not met so that things that indicated the items could be analyzed were not. This is in line with the results of the study which also stated that the results of the evaluation of invalid items were influenced by the statement of the items that were not understood by the test takers. The items that were structured were not objective conditions or the test participants themselves who answered haphazardly valid (Simamora et al., 2021).

2. Reliability

The test reliability coefficient was 0.79 or categorized as high with an average score of 25.57 and a standard deviation of 6.78. The results of the study show that the coefficient was relevant to the reliability coefficient value of 0.79 or the test had a high level of reliability in the sense that the test has a high degree of certainty in assessing what it is assessing and showed a larger index, namely the extent to which the measuring instrument can be trusted (Simamora et al., 2021).

3. Difficulty Level

The test analysis showed that there were 28 difficult items, 20 medium items, and 12 easy items with difficulty indexes ranging from 0.03 to 0.92 with the percentage of difficult, medium, and easy items respectively, namely 47%, 33%, and 20%, the portion was 5:3:2 and the test difficulty index obtained was at 0.35 which was categorized as moderate or sufficient. However, the proportion does not meet the requirements for a significant and good comparison to be given to students. This is relevant to the results of the study which stated that the comparison of the difficulty levels of the items was not good because it was disproportionate which should have obtained a balance whose portion fulfilled 3:4:3 or 3:5:2 (Magdalena et al., 2021).

4. Discriminating Power

The test showed that there were 12 very bad items, 16 bad items, 18 moderate items, 13 good items, and 1 very good item. The discriminating power that is considered sufficient for a question, that is, if it is equal to or greater than 0.30. If it is less than 0.30, then the item is considered to be less able to distinguish test takers who are prepared to face the test from test takers who are not

prepared. Therefore, these items was removed from the test instrument. If the higher the discriminating power of an item, the better the item is. Conversely, if the discriminating power of an item is lower, the item is considered not good.

5. Effectiveness of Distractor

The item distractor worked well and is effective with 13 items whose options were rejected or needed to be replaced and 56 items whose options were revised or needed to be corrected. The effectiveness of the distractor was obtained from the number of test takers who chose options A, B, C, D, and E or did not choose any option so that from the distribution pattern of the answers it could be determined whether the distractor was working or not. A distractor can be said to function if the distractor has great appeal for test takers who do not understand the concept being tested and the distractor has function if the lower the ability level of the test takers the more choose the distractor (Quaigrain & Arhin, 2017).

6. Critical Thinking Skills

Figure 2 shows that of the 37 test takers who participated, 16% had low levels of critical thinking ability, 65% medium, and 19% high. This shows a social phenomenon in everyday life, namely in a certain group, some people are found to have below-average abilities, some have above-average abilities, and the majority have medium or medium abilities. Meanwhile, if the test takers are part of a certain population, it can be said that the level of critical thinking abilities in that group shows that the individuals in that group are superior seeds.

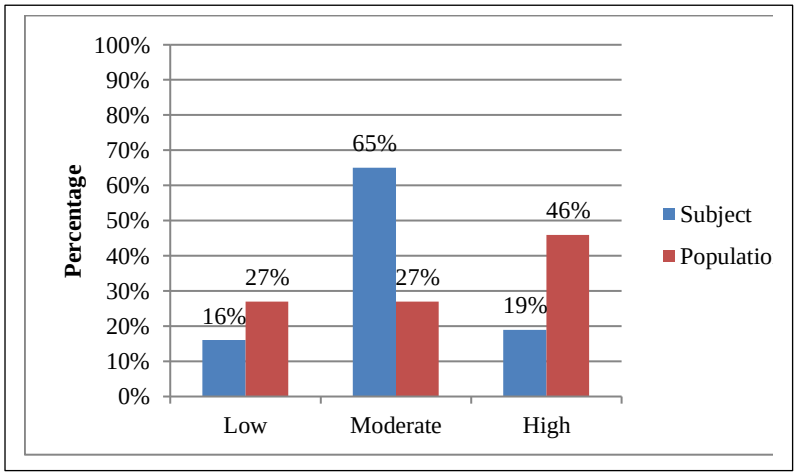


Figure 2. Percentage of students' critical thinking skills

Figure 3 shows that of the 5 indicators of critical thinking skills measured, 20% of test takers mastered interpretation, 17% inference, 21% analysis, 13% evaluation, and 29% self-regulation. Based on these results it can be stated that HOTS questions that measure inference and evaluation skills are difficult to solve compared to HOTS questions that

measure self-regulation skills. The highest mean scores were aspects of self-regulation and the lowest average scores were aspects of evaluating and concluding. The CTS indicators, namely interpretation, inference, analysis, evaluation, and self-regulation measured with the HOTS questions are described as follows.

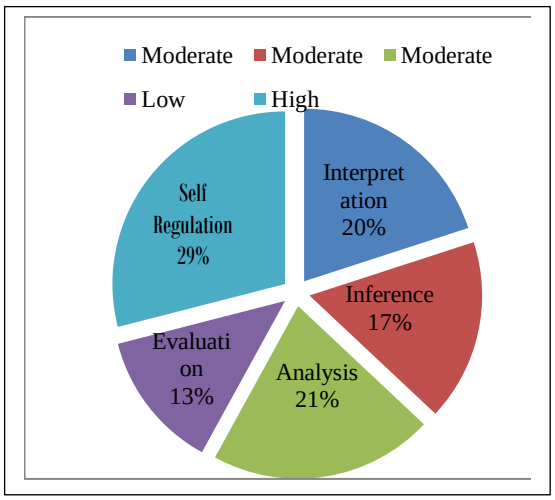


Figure 3. Percentage of students' CTS based on indicators

1. Interpretation

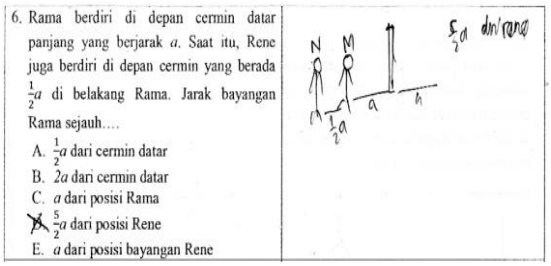

6. Rama berdiri di depan cermin datar panjang yang berjarak *a*. Saat itu, Rene juga berdiri di depan cermin yang berada $\frac{1}{2}a$ di belakang Rama. Jarak bayangan Rama sejauh....
A. $\frac{1}{2}a$ dari cermin datar
B. $2a$ dari cermin datar
C. a dari posisi Rama
D. $\frac{5}{2}a$ dari posisi Rene
E. a dari posisi bayangan Rene

Figure 4. HOTS problems that measure interpretation

Figure 4 shows that the students can paraphrase or change the words in the items in the form of information related to Rama's distance from a flat mirror and Rene's distance from behind Rama so that it becomes a picture form. Next, construct the images of Rama and Rene that are formed behind a flat mirror in order to obtain a solution to the desired problem in the item in the form of the exact distance of Rama's image to the desired specific variable.

2. Inference

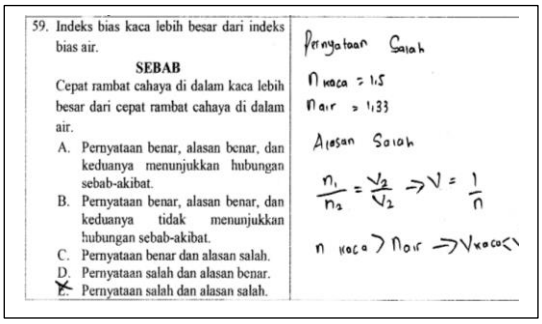

59. Indeks bias kaca lebih besar dari indeks bias air.
SEBAB
Cepat rambat cahaya di dalam kaca lebih besar dari cepat rambat cahaya di dalam air.
A. Pernyataan benar, alasan benar, dan keduanya menunjukkan hubungan sebab-akibat.
B. Pernyataan benar, alasan benar, dan keduanya tidak menunjukkan hubungan sebab-akibat.
C. Pernyataan benar dan alasan salah.
D. Pernyataan salah dan alasan benar.
E. Pernyataan salah dan alasan salah.

Figure 5. HOTS problems that measure inference

Figure 5 shows that test takers can find the fact that the refractive index of glass is 1.5 and the refractive index of water is 1.33. However, the test takers were wrong in concluding that the statement $n_{glass} > n_{air}$ was true. This is an error in critical thinking or an error that occurs in the process of training critical thinking skills, namely drawing conclusions hastily. This can be caused by the decreasing time for completing the test or the amount of time needed to complete the questions one by one.

3.Analysis

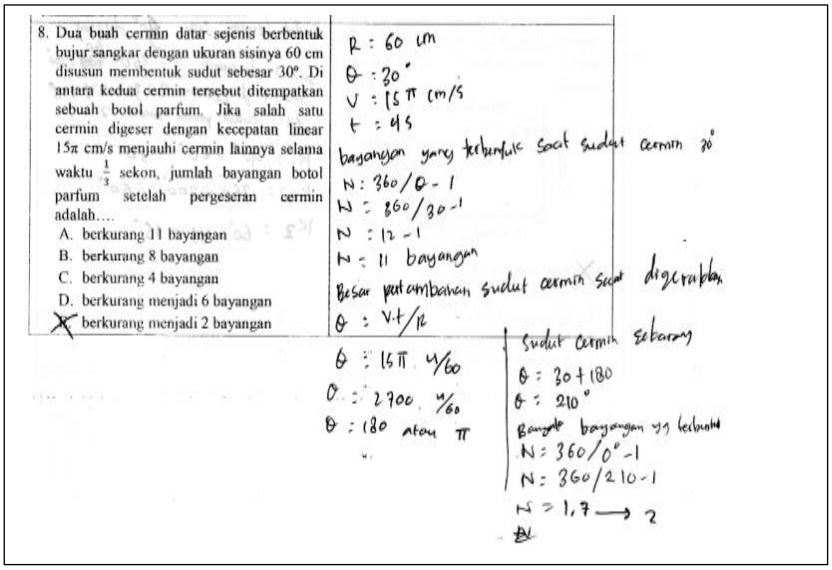


Figure 6. HOTS problems that measure analysis

Figure 6 shows that students can determine the number of images formed when the angle θ (before the mirror is shifted). However, it is wrong to calculate the angle that is shifted away from the mirror so the calculation of the number of images formed (after the mirror is shifted) is also wrong and an incorrect solution is obtained.

4. Evaluation

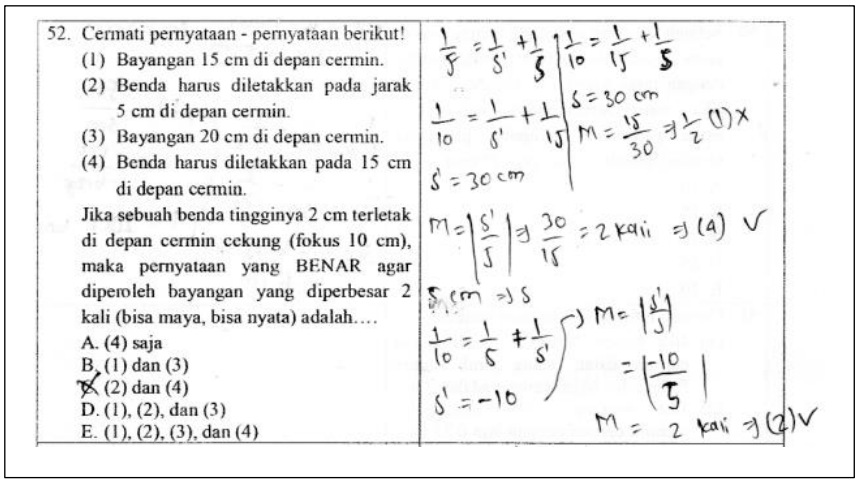
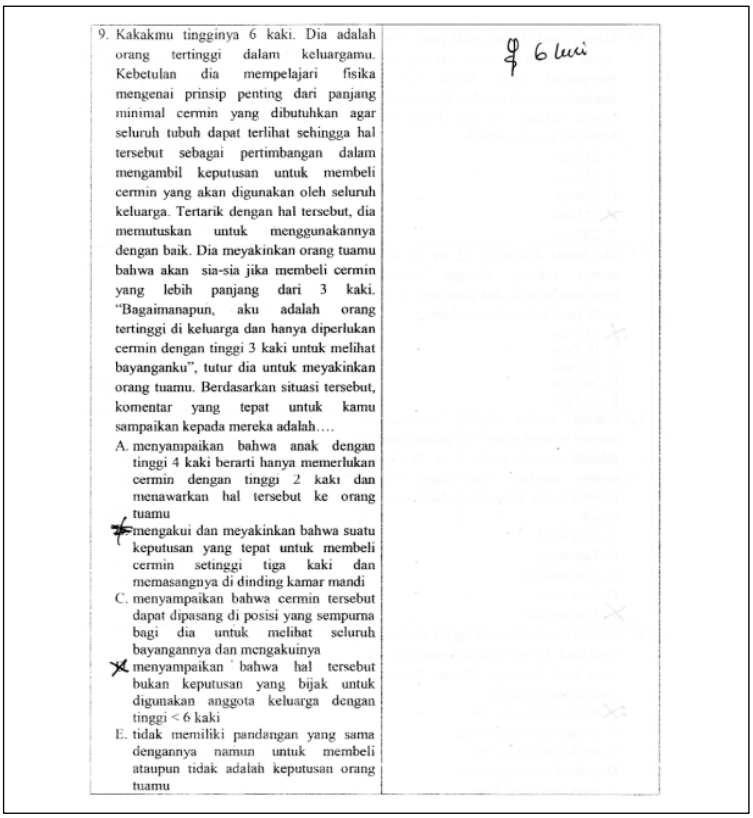


Figure 7. HOTS problems that measure evaluation

Figure 7 shows that the students can solve the problem by checking each statement according to the given conditions. To check each statement, the students needs to apply several concepts regarding image magnification in a concave mirror. Statements (1) and (3) are incorrect because the magnifications are not appropriate.

Meanwhile, statements (2) and (4) are correct because the magnification corresponds to the conditions given, namely twice.

5. Self-Regulation


9. Kakakmu tingginya 6 kaki. Dia adalah orang tertinggi dalam keluargamu. Kebetulan dia mempelajari fisika mengenai prinsip penting dari panjang minimal cermin yang dibutuhkan agar seluruh tubuh dapat terlihat sehingga hal tersebut sebagai pertimbangan dalam mengambil keputusan untuk membeli cermin yang akan digunakan oleh seluruh keluarga. Tertarik dengan hal tersebut, dia memutuskan untuk menggunakannya dengan baik. Dia meyakinkan orang tuamu bahwa akan sia-sia jika membeli cermin yang lebih panjang dari 3 kaki. "Bagaimanapun, aku adalah orang tertinggi di keluarga dan hanya diperlukan cermin dengan tinggi 3 kaki untuk melihat bayanganku", tutur dia untuk meyakinkan orang tuamu. Berdasarkan situasi tersebut, komentar yang tepat untuk kamu sampaikan kepada mereka adalah....

A. menyampaikan bahwa anak dengan tinggi 4 kaki berarti hanya memerlukan cermin dengan tinggi 2 kaki dan menawarkan hal tersebut ke orang tuamu
B. mengakui dan meyakinkan bahwa suatu keputusan yang tepat untuk membeli cermin setinggi tiga kaki dan memasangnya di dinding kamar mandi
C. menyampaikan bahwa cermin tersebut dapat dipasang di posisi yang sempurna bagi dia untuk melihat seluruh bayangannya dan mengakuinya
D. menyampaikan bahwa hal tersebut bukan keputusan yang bijak untuk digunakan anggota keluarga dengan tinggi < 6 kaki
E. tidak memiliki pandangan yang sama dengannya namun untuk membeli ataupun tidak adalah keputusan orang tuamu

Figure 8. HOTS problems that measure self-regulation

Figure 8 shows that the students can be aware that a mirror with a length of 6 feet cannot be used by a family <6 feet if it is installed according to the needs of older siblings and the same thing will happen vice versa if it is installed according to the needs of other family members.

Based on the explanation above, it can be said that validity and reliability can determine whether a product's quality has been improved so that it can be applied as a testing standard or gold product. Meanwhile, the difficulty level, distinguishing power, and effectiveness of distractors can determine the quality of the questions being developed. The quality of the items is said to be sufficient with a difficulty level that is almost close to the proportion it should be, 12 very bad items were rejected, 16 destructive items were revised, and there are 13 items whose options were rejected, 56 items whose options needs to be repaired. In addition, CTS indicators, namely interpretation, inference, analysis, evaluation, and self-regulation can be measured by HOTS problems.

According to Simamora et al. (2021), the results of the study show that a test validity coefficient of 0.72 or categorized as high, and

a test reliability coefficient of 0.79 or categorized as high are relevant with previously conducted research. Besides that, according to Magdalena et al. (2021), the results of the difficulty level test are relevant to previously conducted research which stated that the comparison of the difficulty levels of the items was not good because it was disproportionate which should have obtained a balance whose portion fulfilled 3:4:3 or 3:5:2. This statement holds in line with the results of its discriminating power and deceptive effectiveness. Meanwhile, in integrating HOTS problems to measure indicators of students' critical thinking skills, namely by analyzing more deeply the slices of operational verbs that apply to both so that it is obtained that CTS indicators can be measured with HOTS problems. Relevant to this statement Kahar et al. (2021); Asiah, (2021); Saepuzaman et al (2022), said that high-level cognitive processes are a core element of student CTS.

There are several implications and limitations found in this study. The positive result is that innovations in previous research can be upgraded, especially in terms of the research procedures carried out. Apart from that, teachers have references regarding creating HOTS questions which are used to measure students' critical thinking skills. Therefore, this research can broaden the point of view for future researchers and the school can gain a new mindset to always collaborate in a discussion forum with policy stakeholders and vice versa at the school so that knowledge will always develop. Moreover, although this research discusses the topic of assessment instruments, in terms of procedures it can be used to innovate in the development of learning media, student worksheets, and other learning tools. It is expected that students can get used to working on HOTS questions if teachers are used to and understand how HOTS questions are created and developed so that in the future they can be integrated into their learning process or in the implementation of physics practicum at school.

IV. CONCLUSION AND SUGGESTION

This study produced a HOTS test which was validated conceptually with an Aiken V index of 0.69 in the medium category. The HOTS test was also tested empirically; there were 31 valid questions out of 60 items tested with a product-moment correlation coefficient of 0.72 in the high category and a test reliability coefficient of 0.79 in the high category. The item difficulty index is in the range of 0.03 to 0.92 of which 12 items with low discriminating power need to be revised, while 16 items with very low discriminating power need to be replaced. The effectiveness of the distractor shows that there are 13 questions with options that need to be replaced. Based on trials given to 37 students on the developed HOTS test, critical thinking skills can be described as 16% low, 65% moderate, and 19% high so that the final

product of the HOTS test developed consisted of 32 items and was declared standardized or had met the standard test requirements.

This research has several weaknesses including not paying attention to the proportions of the scope or coverage of the material both in terms of level of difficulty, dimensions of knowledge, and cognitive process dimensions in Bloom's revised taxonomy in designing test grids and also its implementation can be tested on a wide range of students, so it is expected that further researchers can produce even better research products in the future, especially for different physics materials.

ACKNOWLEDGMENTS

Acknowledgments to the teachers, lecturers, and physics education staff at Halu Oleo University who have provided assistance throughout the research process. Likewise, Bimbel 4JO Kendari which has provided financial support for the research, development, and publication of this article.

REFERENCES

Abidin, A. Z., Istiyono, E., Fadilah, N., & Dwandaru, W. S. B. (2019). A computerized adaptive test for measuring the physics critical thinking skills. *International Journal of Evaluation and Research In Education. 8*(3), 376-383. http://doi.org/10.11591/ijere.v8i3.19642

Ali, C. A., Acquah, S., & Esia-Donkoh, K. (2021). A comparative study of SAM and ADDIE models in simulating STEM instruction. *African Educational Research Journal*, *9*(4), 852–859. https://doi.org/10.30918/aerj.94.21.125

Arifin, Z. (2013). *Evaluasi pembelajaran*. PT Remaja Rosdakarya.

Arikunto, S. (2012). *Prosedur penelitian*. Rineka Cipta.

Asiah, N. (2021). The effect of guided inquiry learning models on students' critical thinking skills and learning outcomes in science subjects at MTs Miftahul Muin. *Jurnal Pendidikan Fisika*, 9(2), 165-177. https://doi.org/10.26618/jpf.v9i2.5141

Damayanti, A. E., & Kuswanto, H. (2020). The use of android-assisted comics to enhance students' critical thinking skill. *Journal of Physics: Conference Series*, *1440*, 1-7. doi:10.1088/1742-6596/1440/1/012039

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (PysTHOTS) peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Kahar, M. S., Syahputra, R., Arsyad, R. B., Nursetiawan, N., & Mujiarto, M. (2021). Design of student worksheets oriented to higher order thinking skills (HOTS) in physics learning. *Eurasian Journal of Educational Research, 96*, 14-29.

Khaeruddin, K., Indarwati, S., Sukmawati, S., Hasriana, H., & Afifah, F. (2023). An analysis of students' higher order thinking skills through the project-based learning model on science subject. *Jurnal Pendidikan Fisika Indonesia*, *19*(1), 47–54. https://doi.org/10.15294/jpfi.v19i1.34259

Lespita, E., Purwanto, A., & Syarkowi, A. (2023). Application of problem based learning model assisted by augmented reality media to improve students' high order thinking skills. *Jurnal Pendidikan*

Fisika, 11(1), 1-12. https://doi.org/10.26618/jpf.v11i1.9069

Litna, K. O., Mertasari, N. M. S., & Sudirtha, G. (2021). Pengembangan instrumen tes higher order thinking skills (HOTS) matematika SMA kelas X. *Jurnal Penelitian dan Evaluasi Pendidikan Indonesia*, *11*(1), 10–20. https://doi.org/10.23887/jpepi.v11i1.278

Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis validitas, reliabilitas, tingkat kesulitan, dan daya beda butir soal ujian akhir semester tema 7 kelas III SDN Karet 1 Sepatan. *Bintang : Jurnal Pendidikan dan Sains*, 3(2), 198–214.

Marnah, Y., Suharno., & Sukarmin. (2021). Development of physics module based high order thinking skill (HOTS) to improve student's critical thinking. *Journal of Physics*, *2165*, 1-6. doi 10.1088/1742-6596/2165/1/012018

Nisa, S. K., & Wasis. (2018). Analisis dan pengembangan soal high order thinking skills (HOTS) mata pelajaran fisika tingkat Sekolah Menengah Atas (SMA). *Inovasi Pendidikan Fisika*, *7*(2), 201–207. https://doi.org/10.26740/ipf.v7n2.p%25p

Nurilma, F. R., Supriana, E., & Diantoro, M. (2023). Using STEM-Based 3D-multimedia to improve students' critical thinking skills in uniform circular motion. *Jurnal Pendidikan Fisika*, *11*(2), 193–201. https://doi.org/10.26618/jpf.v11i2.10785

OECD. (2019). *Pendidikan di Indonesia belajar dari hasil PISA 2018.* Pusat Penilaian Pendidikan Balitbang Kemendikbud

Quaigrain, K., & Arhin, A. K. (2017. Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1-11. https://doi.org/10.1080/2331186X.2017.1301013

Redhana, I. W. (2019). Mengembangkan keterampilan abad ke-21 dalam pembelajaran kimia. *Jurnal Inovasi Pendidikan Kimia*, *13*(1), 2239-2253. https://doi.org/10.15294/jipk.v13i1.17824

Saepuzaman, D., Istiyono, E., Haryanto. (2022). Characteristics of fundamental physics higher-order thinking skills test using item response theory analysis. *Pegem Journal of Education and Instruction*, *12*(4), 269-279. https://doi.org/10.47750/pegegog.12.04.28

Saputro, B., Supahar. (2018). Pengembangan instrumen penilaian kemampuan berpikir tingkat tinggi untuk mengukur pencapaian hasil belajar fisika peserta didik SMA kelas XI materi optika. *Jurnal Pendidikan Fisika*, 1–6.

Setiyoningtyas, R., & Kasmui. (2020). Pengembangan quizizz-assisted test berbasis literasi peserta didik pada materi larutan elektrolit nonelektrolit. *Chemistry in Education*, *9*(2), 63–69.

Shaheen, N. (2016). International students' critical thinking–related problem areas: UK university teachers' perspectives. *Journal of Research in International Education*, *15*(1), 18–31. https://doi.org/10.1177/1475240916635895

Sidik, Y., Ishartono, N., Desstya, A., Prayitno, H. J., Anif, S., Hidayat, M. L. (2021). Improving elementary school students' critical thinking skill in science through hots-based science questions: A quasi-experimental study. *Jurnal Pendidikan IPA Indonesia*, *10*(3), 378-386. https://doi.org/10.15294/jpii.v10i3.30891

Simamora, H., Hartono., & Effendi. (2021). Analisis kualitas butir soal buatan guru

kimia pada tes ujian tengah semester ganjil kelas XII MIPA. *Hydrogen: Jurnal Kependidikan Kimia*, 9(1), 8-18. https://doi.org/10.33394/hjkk.v9i1.3701

Sudjana. (2005). *Metode statistika*. Tarsito

Sukmagati, P. O., Yulianti, D., & Sugianto. (2020). Pengembangan lembar kerja siswa (LKS) berbasis STEM (Science, Technology, Engineering, and Mathematics) untuk meningkatkan kemampuan berpikir kreatif siswa SMP. *Unnes Physics Education Journal*, *9*(1), 18–26. https://doi.org/10.15294/upej.v9i1.38277

Widjanarko, P. B. (2022). Penerapan pembelajaran dan penilaian berorientasi higher order thinking skill (HOTS) dalam pelajaran fisika dengan pokok bahasan besaran dan satuan di SMA Charitas Jakarta. *Science : Jurnal Inovasi Pendidikan Matematika dan IPA*, *2*(3), 405–414. https://doi.org/10.51878/science.v2i3.1590