|  |
| --- |
| Volume 12 Nomor 4 (2024), 111- 121**Jurnal Pelita Pendidikan**Journal of Biology Education<https://jurnal.unimed.ac.id/2012/index.php/pelita/index> eISSN: [2502-3217](http://u.lipi.go.id/1452573392) pISSN: [2338-3003](http://u.lipi.go.id/1331887324) |

**DEVELOPMENT OF HOTS-BASED BIOLOGY SUMMATIVE TEST INSTRUMENT AT ODD SEMESTER OF CLASS XI**

**Elsariya Br Ginting1\*, Ely Djulia2**

*1Department of Biology, State University of Medan, Indonesia*

*2Department of Biology, State University of Medan, Indonesia*

\*Corresponding Author: elssariyaginting@gmail.com

**ARTICLE INFO**

**Article History:**

Received: December 02st, 2024

Revised: December 09st, 2024

Accepted: December 12st, 2024

**Keywords:**

*Biology summative test, HOTS, test instrument*

**ABSTRAK**

Penelitian ini bertujuan untuk mengetahui kelayakan instrumen tes sumatif biologi berbasis HOTS berdasarkan aspek validitas, reliabilitas, indeks kesukaran, indeks daya pembeda, dan fungsi distraktor. Penelitian ini dirancang dengan menggunakan model pengembangan ADDIE. Instrumen tes divalidasi oleh para ahli dan melalui tiga kali uji coba. Hasil penelitian menunjukkan bahwa instrumen tes dinyatakan sangat layak berdasarkan ahli materi dan ahli evaluasi. Tiga kali uji coba terhadap soal pilihan ganda menunjukkan: (1) kecenderungan peningkatan validitas (berturut-turut 85%, 95%, 100%), (2) kecenderungan peningkatan reliabilitas (berturut-turut 0.72, 0.75, 0.77) dengan kategori tinggi, (3) indeks kesukaran didominasi oleh soal dengan tingkat kesukaran sedang dan hampir mendekati proporsi yang seimbang, (4) indeks daya pembeda didominasi oleh butir soal berkategori baik, dan (5) distraktor berfungsi secara efektif. Sementara itu, tiga kali uji coba terhadap soal uraian menunjukkan: (1) kecenderungan peningkatan validitas (berturut-turut 80%, 100%, 100%), (2) kecenderungan penurunan reliabilitas (berturut-turut 0.71, 0.71, 0.64) dengan kategori cukup, (3) indeks kesukaran didominasi oleh soal dengan tingkat kesukaran sedang namun masih jauh dari proporsi yang seimbang, dan (4) indeks daya beda didominasi oleh butir soal berkategori baik. Dapat disimpulkan bahwa instrumen tes berbasis HOTS yang dikembangkan layak digunakan sebagai instrumen evaluasi.

**ABSTRACT**

This research aims to determined the feasibility of HOTS-based biology summative test instruments from the aspects of validity, reliability, difficulty index, discriminating index, and distractor function. This research was designed using ADDIE development model. The test instrument was validated by experts and undergo three times trial. Research results showed that the test instrument was declared very feasible according to material expert and evaluation expert. Three times trial of multiple-choice question showed: (1) increased trend of validity (respectively 85%, 95%, 100%), (2) increased trend of reliability (respectively 0.72, 0.75, 0.77) with high category, (3) the difficulty index is dominated by medium levels and nearly approached the balanced proportion, (4) the discriminating index are dominated by good category items, and (5) the distractors are effectively functioning. While three times trial of essay question showed: (1) increased trend of validity (respectively 80%, 100%, 100%), (2) decreased trend of reliability (respectively 0.71, 0.71, 0.64) with enough category, (3) the difficulty index is dominated by medium levels but it was still far from the balanced proportion, and (4) the discriminating index are dominated by good category items. It can be concluded that the HOTS-based test instrument is feasible to be use as an evaluation instrument.

Copyright © 2024 Universitas Negeri Medan. Artikel Open Access dibawah

lisensi CC-BY-4.0 (https://creativecommons.org/licenses/by/4.0)

|  |
| --- |
| **How to Cite:**Ginting, B.E., & Djulia, Ely. (2024). Development of Hots-Based Biology Summative Test Instrument at Odd Semester of Class XI, 12(4): 111-121. |

**INTRODUCTION**

The paradigm of education in 21st-century is accompanied by rapid technological developments that provide new challenges and complex problems that students must be faced with. In order to be able to answer the challenges of the era, student must be prepare to be trained in critical thinking, creativity, and problem solving (Sani, 2019). These three abilities are referred to as higher-order thinking skills (HOTS).

HOTS requires a person to understand, infer, connect factual information with theoretical concepts, categorize, manipulate, find facts within a given context, and propose a solution to a problem that arises (Thomas and Thorne, 2009). As defined by Retnawati (2018), HOTS is an incision between the three primary components of the cognitive process dimension (analysis, evaluation, and, creation) and the three primary components of the knowledge dimension (conceptual, procedural, and metacognitive).

Based on the results of the Program Internationale for Student Assessment (PISA) in 2022, Indonesia is ranked 69th out of 81 countries that participated in this test (OECD, 2023). The questions used in PISA are the type of questions that assess students' ability to solve contextual life problems and students' ability to think at a higher level. However, the low rank of Indonesia in this study indicates that students' ability to think at a higher level is still low. One of the factors is because Indonesian students are less trained to solve high-level thinking questions due to the lack of HOTS-based questions used in the school (Junaidi et al., 2020). This is in line with the research result by Astuti (2021) which stated that students' HOTS is still low and the improvement is needed. Students' HOTS achievement is strongly influenced by the learning experienced by the students.

One of the efforts made by the government to improve and train students' HOTS is by the implementation of 2013 curriculum which focuses on the curriculum improvement in two major parts, namely content standards and assessment standards. Improvement in the assessment standards is carried out by adapting international standard assessment models, which place greater emphasis on HOTS-based evaluation system (Kemendikbud, 2017). The 2013 curriculum requires student to demonstrate the ability to estimate, plan, and predict. This is in line with the realm of HOTS, which involves analysis (the ability to think in specializing certain aspects); evaluation (the ability to think in making decisions based on contextual life problems); and creating (the ability to think in constructing broad insights owned by the students) (Umami et al., 2021).

The HOTS can be viewed from two perspectives, according to Bloom taxonomy and Brookhart's. In Bloom's taxonomy, the achievement of high-level thinking skills is included in the C4, C5, and C6 categories (Anderson and Krathwohl, 2001). Meanwhile, according to Brookhart (2010), indicators of HOTS problems include critical thinking, problem solving or solution finding, and creative thinking.

In order to train and measure students' HOTS, teachers can administer HOTS-based assesment as an alternative strategy (Kusuma, 2017). However, the main problem found in school is the lack of teachers understanding in preparing and developing HOTS questions, such as the skills to present various contextual stimulus in the form of text, images, graphs, and tables containing problems related to student’ daily life (Merta et al., 2019), therefore the availability of HOTS questions in schools is still lacking and the use of HOTS-based test instrument is still rarely used in school assessment system.

Due to the problems encountered in the field, namely the lack of HOTS-based test instrument while curriculum demands the HOTS-based learning assessments, therefore the development of HOTS-based test instrument is important to do.

In developing HOTS-based test, there are several steps must be followed: (1) analyzing basic competencies that can be made into HOTS questions; (2) compiling a questions grid; (3) selecting the contextual stimulus; (4) determining the answer key; and (5) performing item analysis (Widana, 2017). This can be done by analyzing the item based on the aspect of validity, reliability, difficulty index, discriminating index, and effectiveness of distractor function. Globally, according to Adams and Wieman (2011), research and development of evaluation go through several stages, including: (1) mapping the test objectives and the scope of the construct or domain level to be measured; (2) development and evaluation of specification tests; (3) development, field testing, evaluation, item selection and design scoring guidelines, and (4) construction and evaluation of tests for operational use.

The development of HOTS-based assessment is important to do due to the fact that the quality of assessment greatly affects the quality of learning progress as a whole. According to Airasian (1994), assessment has three main purposes, namely to assist learning, measure student achievement, and evaluate the entire program. Assessment techniques consist of two types, namely test instruments and non-test instruments. Test instruments are classified into objective and subjective tests.

Many research related to the development of HOTS assessment instrument has been carried out, such as the research conducted by Fidia et al. (2022) which succeeded in developing a valid and reliable HOTS question instrument on plants tissue and organ material. The result of theory validity is very valid 90,6% in material aspect, 90% in construction aspect, 100% in language aspect. The result of emphiric validity belongs to valid 80%. The reliability is 0,818. Difficulty index is 70% medium. The discriminating index is 40% good. As well as research by Ulfa and Kuswanti (2021) that showed the developed HOTS assessment instrument of the respiratory system that was valid and reliable with proportion of difficulty index was 60% moderate and 40% difficult.

This paper describes the development of a HOTS-based biology summative test instrument for class XI at odd semester. Researcher aims to develop HOTS-based test instrument that is feasible in the aspects of validity, reliability, difficulty index, discriminating index, and distractor function.

**METHOD**

This research and development (R&D) used the ADDIE development model by Dick and Carry (1996) that consists of 5 stages, namely analysis, design, development, implementation, and evaluation. The test instrument was developed based on the guidelines for writing HOTS test by the Ministry of Education and Culture in the 2013 curriculum. This research was carried out in SMAN 2 Medan. The subjects of this study are grade XI MIA students, consists of 34 students for the first trial, 90 students for the second trial, and 165 students for the third trial. Data collection technique were in the form of validation sheet by expert and test.

The development of HOTS-based test instrument going through 5 stages, namely:

1. **Analysis Stage**

The focus of this stage is to collecting the initial research informations through needs analysis, curriculum analysis, and analysis of previous test instrument that have been used in school last year.

1. **Design Stage**

The focus is on the following steps for preparing HOTS questions, as follows: (1) determining the basic competencies and materials to be assessed; (2) arranging the question grids; (3) developing the draft of HOTS-based test instrument; (4) designing the scoring guidelines/answer key; and (5) modifying expert validation sheets. In this stage, the material specified is cell, plant tissue, animal tissue, human movement system, and human circulatory system. After that, a grid of questions, scoring guidelines, and expert validation sheets was made. At the end of this stage, an initial design totaling of 25 questions was obtained, consist of 20 items of multiple-choice and 5 items of essay questions. The proportion of question is 13 items of C4, 9 items of C5, and 3 items of C6.

1. **Development Stage**

In this stage, instrument feasibility test was carried out using expert validation sheet. HOTS-based test instrument that have been developed as an initial product in the previous stage must be evaluated by qualified validators including material expert and evaluation expert. All of the suggestion and comments from the experts was used as guidelines for the revision and improvement of the test instrument until it declared valid. The assessment of validator follows a Likert scale of 1-4 for each assessment item. The average assessment results then classified according to the criteria presented in table 1. Data was analyzed based on following formula:

*x̄ =* $\frac{Σx}{n}$ *× 100%*

Description:

x̄ = The average value of instrument validity

Σx = The total of empiric validator score

n = Maximum score

Table 1. Evaluation Criteria for Instrument Validity

|  |  |
| --- | --- |
| Average Value | Evaluation Criteria |
| 25% - 40% | Invalid (may not be used) |
| 41% - 55% | Less valid (may not be used) |
| 56% - 70% | Valid enough (may be used after major revision) |
| 71% - 85% | Valid (may be used after minor revision) |
| 86% - 100% | Very valid (very good used) |

1. **Implementation Stage**

In this stage, product trial was conducted. The HOTS-based test instrument undergoes three times trial involving the research subjects. The result of this stage was used to analyzed and determined the feasibility of the test instrument that have been developed based on the aspect of validity, reliability, difficulty index, discriminating index, and distractor function.

1. **First trial**

The first trial was conducted by administered 25 test items into 34 students of class XI MIA 2 within 90 minutes. The results of this stage showed that from 25 items there are 17 items of multiple-choices and 4 items of essay that meet all of the analysis criteria. However, there are 3 items of multiple-choices and 1 item of essay that did not meet the analysis criteria. 2 items were revised and the other 2 items were discarded due to the bad discriminating index. After revised, there were a total of 23 items (19 items of multiple-choice and 4 items of essay) that can be retested in the second trial.

1. **Second trial**

The second trial was conducted by administered 23 test items into 90 students of class XI MIA 1, XI MIA 3, and XI MIA 4 within 90 minutes. The results of this stage showed that from 23 items, there are 18 items of multiple-choices and 4 items of essay that meet all of the analysis criteria. There is 1 item of multiple-choice question that did not meet the analysis criteria, that item was discarded due to the bad discriminating index. Finally, there were a total of 22 question items (18 items of multiple-choice and 4 items of essay) that can be retested in the third trial.

1. **Third trial**

The third trial was conducted by administered 22 test items into 165 students of class XI MIA 5, XI MIA 6, XI MIA 7, XI MIA 8, and XI MIA 9 within 90 minutes. The results of this stage showed that all of the items already meet all of the analysis criteria. From this stage, the final result was obtained. There are 22 question items (18 items of multiple-choice and 4 items of essay) that are feasible in the aspects of validity, reliability, difficulty index, discriminating index, and distractor function.

1. **Evaluation Stage**

This is the last stage of this research. In this stage researchers analyzed the result of implementation stage. The test item was analyzed based on the validity, reliability, difficulty index, discriminating index, and distractor function. The analysis was done using Microsoft Excel program.

1. **Validity**

To calculate the validity of multiple-choice question items, the following formula can be use:

$r\_{pbi}$ *=* $\frac{M\_{p}-M\_{t}}{SD\_{t}}\sqrt{\frac{P}{q}}$

Description:

$r\_{pbi}$= Biserial correlation coefficient

$M\_{p}$ = The average score of the students who

 answer correctly

$M\_{t}$ = The average total score

$SD\_{t}$ = Standard deviation of total score

p = The proportion of students who answered

correctly

q = The proportion of students who answered

incorrectly

To calculate the validity of essay question items, the following formula can be use:

*rxy =* $\frac{n(∑XY) - (∑X)(∑Y)}{\sqrt{ (n(∑X^{2}} - (∑X)^{2 })(n(∑Y^{2 }- (∑Y)^{2 })}$

Description:

rxy = Correlation coefficient between X and Y

ΣX = Number of item scores

ΣX = Total score (all items)

n = Number of respondents

The validity value was obtained by comparing the calculated correlation coefficient values with the correlation coefficient table. The benchmark to interpreting the validity of the instrument is as listed in the table 2. Test items can be said to be valid if tcount > ttable (Arikunto, 2017).

Table 2. Correlation Coefficient/ Validity Criteria

|  |  |
| --- | --- |
| Range | Criteria |
| 0,81 – 1,00 | Very high |
| 0,61 – 0,80 | High |
| 0,41 – 0,60 | Medium |
| 0,21 – 0,40 | Low |
| 0,00 – 0,20 | Very low |

1. **Reliability**

To calculate the reliability of multiple-choice questions, KR-20 formula can be use:

$r\_{11}$ ***=*** $\left(\frac{n}{n-1}\right)\left(\frac{S^{2} - Σpq}{S^{2}}\right)$

Description:

$r\_{11}$ = Reliability value

p = Number of subjects who answered correctly

q = Number of subjects who answered incorrectly

Σpq = The sum of p and q

n = Number of items

S = Standard deviation from the test

To calculate the reliability of essay questions, Cronbach Alpha formula can be use:

$r\_{11}$*=*$\left(\frac{k}{k-1}\right)\left(1-\frac{ΣS\_{i}}{S\_{t}}\right)$

Description:

$r\_{11} $= Reliability value

∑Si = The sum of the variances of each item scores

St = Total variance

k = Number of items

The benchmark to interpreting the correlation value (r11) obtained by comparing the reliability coefficient with the value of the question reliability criteria in table 3.

Table 3. Reliability Criteria

|  |  |
| --- | --- |
| Reliability Coefficient | Criteria |
| 0,91 ≤ r11 ≤ 1,00 | Very high |
| 0,71 ≤ r11 ≤ 0,90 | High |
| 0,41 ≤ r11 ≤ 0,70 | Enough |
| 0,21 ≤ r11 ≤ 0,40 | Low |
| 0,00 ≤ r11 ≤ 0,20 | Very low |

1. **Difficulty Index**

To calculate the difficulty index of an item, the following formula can be use:

*P =* $\frac{B}{JS}$

Description:

P = Difficulty index

B = Number of students who answered correctly

JS = Total number of test takers

The benchmark to interpreting the difficulty index is listed in the table 4.

Table 4. Difficulty Index Category

|  |  |
| --- | --- |
| Correlation Coefficient | Category |
| 0,71 ≤ P ≤ 1,00 | Easy |
| 0,31 ≤ P ≤ 0,70 | Medium |
| 0,00 ≤ P ≤ 0,30 | Difficult  |

1. **Discriminating Index**

To calculate the discriminating index of an item, the following formula can be use:

*D =*$\frac{B\_{A}}{J\_{A}} - \frac{B\_{B}}{J\_{B}} $

Description:

D = Discriminating index

$J\_{A}$ = Number of upper group participants

$J\_{B}$ = Number of participants in the lower group

$B\_{A}$= Number of upper group participants who

 answered correctly

$B\_{B}$ = Number of lower group participants who

 answered correctly

The discriminating index can be calculated by grouping all test takers into the upper and lower group. It is done by dividing the entire group of test takers into two equals, 27% of the upper group and 27% of the lower group. The benchmark to interpreting the discriminating index of an item is listed in the table 5.

Table 5. Discriminating Index Criteria

|  |  |
| --- | --- |
| Discriminating Coefficient | Criteria |
| 0,71 ≤ D ≤ 1,00 | Excellent  |
| 0,41 ≤ D ≤ 0,70 | Good  |
| 0,21 ≤ D ≤ 0,40 | Enough  |
| 0,01 ≤ D ≤ 0,20 | Bad |
| 0 < D  | Very Bad |

1. **Distractor Function**

To calculate the distractor function of multiple-choice items, the following formula can be use:

*IP =*$\frac{P}{N} x 100\%$

Description:

IP = Distractor index

P = Number of students who chose distractor

N = Number of students taking the test

The distractor is said to be functioning well if more than 5% of test takers have been chose it. The benchmark to interpreting the distractor function is listed in the table 6.

Table 6. Distractor Function Criteria

|  |  |
| --- | --- |
| Total of Functioning Distractor | Criteria |
| 4 | Very good |
| 3 | Good |
| 2 | Enough |
| 1 | Bad |
| 0 | Very bad |

**RESULT AND DISCUSSION**

**Feasibility of HOTS-based Test Instrument based on Material Expert**

The HOTS-based test instrument was validated by material expert, who are the biology lecturers in Universitas Negeri Medan. Validation was carried out using validation sheets, which evaluated from three aspects including material, construction, and language.

The result of validation by material expert showed that among 25 items of questions that have been developed, 20 multiple-choice items obtained an average score of 100%, respectively for the material, construction, and language aspects. Meanwhile, from 5 essay items it was obtained an average score of 100% respectively for the material, construction, and language aspects. This result was obtained after gone through some revision according to the suggestion given by material expert. The revisions given by material expert were dominated by suggestions for the improvement of construction and language aspect.

Figure 1. Material Expert Validation Result

From the result analysis, it can be concluded that 25 questions that have been developed were meet very valid criteria (100%), therefore the test instrument declared to be feasible to use and worth to trial.

**Feasibility of HOTS-based Test Instrument based on Evaluation Expert**

The HOTS-based test instrument was validated by evaluation expert, who are the biology lecturers in Universitas Negeri Medan. Validation was carried out using validation sheets, which evaluated from four aspects including material, construction, language, and HOTS aspects.

The result of validation by evaluation expert showed that among 25 items of questions that have been developed, 20 multiple-choice items obtained an average score of 90% in material aspect, 89% in construction aspect, 86% in language aspect, and 85% in HOTS aspect. Meanwhile, from 5 essay questions it was obtained an average score of 87% in material aspect, 88% in construction aspect, 88% in language aspect, and 93% in HOTS aspect. This result was obtained after gone through some revision according to the suggestion given by evaluation expert. The revisions given by evaluation expert were dominated by suggestions for the improvement of construction and language aspect.

From the result analysis, it can be concluded that 25 questions that have been developed were meet very valid criteria (88%), therefore the test instrument declared to be feasible to use and worth to trial. Evaluation expert validation results interpreted as figure 2.

Figure 2. Evaluation Expert Validation Result

**Validity Trend of HOTS-based Test Instrument**

Validity testing of items in the HOTS-based test instrument was done using the point biserial formula with the help of Microsoft Excel program. The results of the calculation (rcount) are then consulted with the rtable at the 5% significance level. Test items can be said to be valid if tcount > ttable (Arikunto, 2017). The number of research subjects in the first trial was 34 students, so at the 5% significance level the rtable value was 0,338. Research subjects in the second trial was 90 students, so the rtable value was 0,207. Research subjects in the third trial was 165 students, so the rtable value was 0,152.

Table 7. Recapitulation of MCQ Validity

|  |  |  |
| --- | --- | --- |
| Trial  | Valid  | Invalid  |
| 1st | 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 17, 18, 19, 20 | 10, 15, 16 |
| Total  | 17 (85%) | 3 (15%) |
| 2nd  | 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 12, 13, 14, 15, 16 17, 18, 19 | 1 |
| Total  | 18 (95%) | 1 (5%) |
| 3rd  | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 |  |
| Total  | 18 (100%) | 0 (0%) |

Table 8. Recapitulation of Essay Validity

|  |  |  |
| --- | --- | --- |
| Trial  | Valid  | Invalid  |
| 1st | 1, 3, 4, 5 | 2 |
| Total  | 4 (80%) | 1 (20%) |
| 2nd  | 1, 2, 3, 4 |  |
| Total  | 4 (100%) | 0 (0%) |
| 3rd  | 1, 2, 3, 4 |  |
| Total  | 4 (100%) | 0 (0%) |

Based on three times trial of multiple-choice question, there is increased trend of item categorized as valid (respectively 85%, 95%, 100% for first trial, second trial, and third trial). However, there is decreased trend of item categorized as invalid (respectively 15%, 5%, 0% for first trial, second trial, and third trial). Based on three times trial on essay question, there is increased trend of item categorized as valid (respectively 80%, 100%, 100% for first trial, second trial, and third trial). However, there is decreased trend of item categorized as invalid (respectively 20%, 0%, 0% for first trial, second trial, and third trial).

The results of validity analysis on multiple-choice and essay questions indicate that the questions have performed its function, namely to measure what should be measured, this is in accordance with the theory stated by Arikunto (2017) that a test is considered to be valid if it can measure what is intended to be measured.

The existence of invalid item in the first and second trial can be caused by several factors as stated by Gronlund (1990), that there are three factors influencing the validity of test results, namely evaluation instrument factors, evaluation administration and scoring factors, and students' answers factors. In this HOTS-based test instument trial, the student’ answer factor was believed to present greater influence more than two other factors. This factor includes the tendency of learners to answer quickly, but incorrectly and the desire to try or answer by guessing. This is in line with the theory stated by Gronlund (1990) that in actual practice, student’ answer factor is the most influential factor on item validity.

As follow up to the validity testing, an invalid item should be revised whereas the valid items can be reused. Item that has been declared valid must be maintained by documenting this item into the question bank. Items that declared as invalid must be revised by increasing the technical expertise of researchers in the item preparation (Prajoko et al., 2022).

**Reliability Trend of HOTS-based Test Instrument**

A valid test instrument must be reliable. Reliability testing of items in the HOTS-based test instrument was done using KR-20 formula for multiple-choice question (MCQ), while for essay question, Cronbach Alpha formula was used. The analysis was processed using Microsoft Excel program.

Table 9. Recapitulation of MCQ Reliability

|  |  |  |  |
| --- | --- | --- | --- |
| Trial  | Total Items | Reliability  | Criteria  |
| 1st | 20 | 0.72 | High |
| 2nd  | 19 | 0.75 | High |
| 3rd  | 18 | 0.77 | High |

Table 10. Recapitulation of Essay Reliability

|  |  |  |  |
| --- | --- | --- | --- |
| Trial  | Total Items | Reliability  | Criteria  |
| 1st | 5 | 0,71 | High  |
| 2nd  | 4 | 0,71 | High |
| 3rd  | 4 | 0,64 | Enough |

Based on three times trial of multiple-choice question, there is increased trend of reliability coeffisien (respectively 0.72, 0.75, 0.77 for first trial, second trial, and third trial). While in the essay questions there is decreased trend of reliability coeffisien (respectively 0.71, 0.71, 0.64 for first trial, second trial, and third trial). The first and second trial of essay reliability test showed a stable trend, however there is decreased trend of reliability coeffisien in the third trial. From those result, it can be concluded that multiple-choice question has a high reliability and essay questions has sufficient (enough) reliability.

Generally, the HOTS-based test instrument that have been developed has a high reliability, it means that the instrument showed the good consistency. An instrument is considered to be reliable if it gives fixed result when tested many times. In evaluation terms, the “stability”, “consistency” or “fixity” does not mean that it should always be “the same”, but it should follow the changes consistently. As an example, if the condition of A is initially lower than B, then if a re-measurement is made, A is also lower than B. What is said to be stable or consistent is the same position of a person in a group (Arikunto, 2017).

Regarding the differences reliability test results between the first, second, and third trial, there are some factors may be affecting, one of them is the decrease of total items and the increase of total samples at each trial stage. This is in line with the theory of Gronlund (1990) which states that some factors affecting reliability including the length of test. There is a tendency that the longer a test, the higher the reliability of a test will be. Because the more questions, the more samples will be assessed and the proportion of correct answers will be greater, so the guessing factor will be smaller. Another factors that affecting the reliability is spread of scores, difficulty index, and objectivity.

**Difficulty Index Trend of HOTS-based Test Instrument**

The difficulty index of an items was analyzed to find out whether the question is easy, medium, or difficult category. The data was processed using Microsoft Excel program.

Table 11. Recapitulation of MCQ Difficulty Index

| Trial  | Easy  | Medium  | Difficult  |
| --- | --- | --- | --- |
| 1st | 11, 14 | 1, 2, 4, 5, 6, 7, 8, 9, 10, 13, 15, 16, 17, 18, 19, 20 | 3, 12 |
| Total  | 2 (10%) | 16 (80%) | 2 (10%) |
| 2nd  | 11, 13 | 2, 4, 5, 6, 7, 8, 9, 10, 14, 15, 16, 17, 18, 19 | 1, 3, 12 |
| Total  | 2 (10%) | 14 (74%) | 3 (16) |
| 3rd  | 10, 12 | 1, 3, 4, 5, 6, 7, 8, 9, 13, 15, 16, 17, 18 | 2, 11,14 |
| Total  | 2 (11%) | 13 (72%) | 3 (17%) |

Table 12. Recapitulation of Essay Difficulty Index

|  |  |  |  |
| --- | --- | --- | --- |
| Trial  | Easy  | Medium  | Difficult  |
| 1st | 5 | 1, 2, 3, 4 |  |
| Total  | 1 (20%) | 4 (80%) | 0 (0%) |
| 2nd  |  | 1, 2, 3, 4 |  |
| Total  | 0 (0%) | 4 (100%) | 0 (0%) |
| 3rd  | 3 | 1, 2, 4 |  |
| Total  | 1 (25%) | 3 (75%) | 0 (0%) |

Based on three times trial of multiple-choice difficulty index, there are increased trend of item categorized as difficult (respectively 10%, 16%, 17% for first trial, second trial, and third trial) and stable trend of item categorized as easy (respectively 10%, 10%, 11% for first trial, second trial, and third trial). However, there is decreased trend of item categorized as medium (respectively 80%, 74%, 72% for first trial, second trial, and third trial). While in the essay question there are unstable trend of item categorized as easy (respectively 20%, 0%, 25% for first trial, second trial, and third trial) and also unstable trend of item categorized as medium (respectively 80%, 100%, 75% for first trial, second trial, and third trial).

The results of difficulty index analysis showed that the HOTS-based test instrument that have been developed are dominated by items in medium difficulty levels, it indicates that the instrument is classified as a good in difficulty level terms. This is in accordance with Arikunto (2017) that stated a good question is one that is neither too easy nor too difficult. It means that questions with a medium difficulty index (0.3 < P ≤ 0.7) are good question criteria. However, it does not mean that questions that are too easy or too difficult can not be used, because each level of questions have their own function. A test instrument is said to be good if it has a balanced difficulty level (proportional). The balanced proportion according to Arifin (2011) is as follows: (a) 25% difficult, 50% medium, 25% easy, or (b) 20% difficult, 60% medium, 20% easy, or (c) 15% difficult, 70% medium, 15% easy. From the analysis results of the third trial, it can be concluded that the difficulty index of multiple-choice questions nearly approached the balanced proportion, while the essay question was still far from the balanced proportion.

**Discriminating Index Trend of HOTS-based Test Instrument**

 The analysis of item discriminating index was carried out to determine the level of items ability in distinguishing high ability student from low ability student. The data was processed using Microsoft Excel program.

 In the first trial, among 20 items of MCQ, there are 1 item in the excellent category, 9 items in the good category, 9 items in the enough category, and 1 item in the bad category. One item in the bad category (question item number 16) was discarded because the discriminating index value was zero. In the second trial, among 19 items, there are 3 items in the excellent category, 10 items in the good category, 5 items in the enough category, and 1 item in the bad category. One item in the bad category (question item number 1) was discarded because the discriminating index value was zero. In the third trial, out of 18 items, there are 3 items in the excellent category, 13 items in the good category, 2 items in the enough category, and no item in the bad category.

 In the first trial, among 5 items of essay questions, there are 1 item in the excellent category, 2 items in the good category, 1 item in the enough category, and 1 item in the bad category. One item in the bad category (question item number 2) was discarded. In the second trial, among 4 items of essay questions, there are 2 items in the excellent category, 2 items in the good category, and no item in the enough, bad, or very bad category. In the third trial, among 4 items of essay, all of the items categorized as good.

 After the discriminating index test is carried out, not all items can be used, except for items with discriminating index in the "enough", "good", and "excellent" categories (Arikunto, 2017). Questions that have negative (<0) and zero (<20) of discriminating index should be closely examined to identify if it has been miskeyed or any other problems with it (Airasian, 1994). While according to Zainuri (2021), that kind of item can be discarded right away because the questions cannot distinguish students who have high and low abilities, while the remaining items can be revised or discarded. In this trial stage, there are total of 2 items of multiple-choice questions and 1 items of essay questions that has been discarded due to the bad category of item discriminating index.

The results of discriminating index analysis showed that HOTS-based test instrument that have been developed are dominated by items in the good category. It means that the questions are able to distinguished student abilities. Therefore, the question is proven to be answered correctly by high abilities students and answered incorrectly by low abilities students. This is in line with Arikunto (2017) which state that a good question is one that can be answered correctly by students from the upper group only (high ability students).

Discriminating index indicates how an individual item fares with students who scored high and low on the overall test (Airasian, 1994). If a question can be answered correctly by both high and low abilities students, then the question is not good because it has no discriminating index. Similarly, if all students, both high and low abilities students answered incorrectly, then the question is not good either because it indicates that the question has no discriminating index.

Figure 3. Diagram of MCQ Discriminating Index

Figure 4. Diagram of Essay Discriminating Index

**Distractor Function Trend of HOTS-based Test Instrument**

The distractor function of multiple-choice items needs to be analyzed to find out the effectiveness of the distractor in deceiving the test takers. Multiple-choice question that has been developed has 5 answer choices. The analysis of distractor function was carried out through consideration to the answer distribution pattern by counting the number of students who choose answer choices whether A, B, C, D, E, or do not choose any option (omit).

Table 13. Recapitulation of Distractor Function

|  |  |  |  |
| --- | --- | --- | --- |
| Trial  | Criteria  | Total  | Total  |
| 1st | Very good | 1, 3, 10, 12, 14, 15 | 30% |
|  | Good | 2, 4, 5, 8, 9, 11, 16, 17, 18, 19, 20 | 55% |
|  | Enough  | 6, 7, 13 | 15% |
|  | Bad |  | 0% |
|  | Very bad |  | 0% |
| 2nd  | Very good | 2, 3, 6, 7, 12, 15, 17 | 37% |
|  | Good | 1, 4, 8, 9, 10 11, 13, 14, 16, 18, 19 | 58% |
|  | Enough  | 5 | 5% |
|  | Bad |  | 0% |
|  | Very bad |  | 0% |
| 3rd  | Very good | 1, 2, 3, 5, 6, 7, 9, 11, 12, 13, 15, 16 | 67% |
|  | Good | 8, 10, 14, 18 | 22% |
|  | Enough  | 4, 17 | 11% |
|  | Bad |  | 0% |
|  | Very bad |  | 0% |

 Analysis result of multiple-choice question (MCQ) in the HOTS-based test instrument showed several answer patterns. Each patterns indicates different possible reason. There are several answer patterns according to Airasian (1994), the first answer pattern is typical of MCQ items that have two correct answers. In this case, the final decision about whether the item or the students are at fault rests with the teacher. The second pattern is one where most studens choose an option other than the answer key. In this case, closer inspection of the option other that answer key must be done to provide clues as to why the option was chosen so often. It can also occur because of the mismatch between the intentions of the teacher and the understanding captured by the majority of students, so the solution is that the teacher must repeat the teaching so that the students' mastery of the material is match with the teacher's actual intentions (Arikunto, 2017).

 The results of distractor function analysis showed that HOTS-based test instrument that have been developed are dominated by items in the very good category, it indicates that most of the questions have a distractor with great attraction to be chosen as the correct answer for more than 5% of test takers who do not master the material. It means the overall questions distractor is effectives in deceiving the test takers. This is in accordance with Arikunto (2017), that stated a good question is one in which the distractors have been selected by at least 5% of the test takers and omitted by less than 10% of the test takers. A distractor can be said to be properly functioning if the distractor has a great attraction as the right answer for student from lower group.

CONCLUSION

Based on the research results and data analysis, several point can be concluded, namely; (1) the test instrument is declared valid and very feasible to use based on material experts and evaluation experts, (2) Three times trial of multiple choice questions showed an increasing trend in validity (respectively 85%, 95%, 100%), reliability was high, the difficulty index is dominated by medium level and nearly approached the balanced proportion, the discriminating index is dominated by items of good category, and distractors are effectively functioning. (3) Three times trial of essay questions showed an increasing trend in validity (respectively 80%, 100%, 100%), reliability was enough, the difficulty index was dominated by medium level but still far from the balanced proportion, and the discriminating index was dominated by items of good category. Therefore, the HOTS-based biology summative test instrument that have been developed is feasible to be use as biology evaluation instrument.

REFERENCES

Adams, W. K., & Wieman, C. E. (2011). Development and Validation of Instruments to Measure Learning of Expert-Like Thinking. International Journal of Science Education, 33(9), 1289-1312.

Airasian, P. W. (1994). Classroom Assessment. New York: McGraw Hill.

Anderson, L., & Krathwohl, D. R. (2001). A Taxonomy of Learning, Teaching, and Assessing. A Revision of Bloom’s Taxonomy of Educational Objectives. New York: Longman.

Arifin, Z. (2011). Evaluasi Pembelajaran. Bandung: PT. Remaja Rosdakarya.

Arikunto, S. (2017). Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.

Astuti, F.O., Juhanda, A., & Suhendar. (2021). Analisis Higher Order Thingking Skill (HOTS) Siswa Kelas IX Pada Materi Bioteknologi. Jurnal Pelita Pendidikan, 9(2), 071-077.

Brookhart, S. M. (2010). Assess Higher Order Thinking Skills in Your Classroom. Alexandria: ASCD Publisher.

Fidia, F., Puspitawati, R. P., & Yakub, P. (2022). Pengembangan Instrumen Soal Higher Order Thinking Skills (HOTs) Materi Jaringan dan Organ pada Tumbuhan Kelas XI SMA. Berkala Ilmiah Pendidikan Biologi (BioEdu), 11(3), 745-754.

Gronlund, N. E., & Linn, L. R. (1990). Measurement and Evaluation in Teaching (6th ed.). New York: Macmillan Publishing Company.

Junaidi, J., Roza, Y., & Maimunah, M. (2020). Kemampuan Berpikir Siswa dalam Menyelesaikan Soal HOTs pada Materi Pola dan Barisan Bilangan. Jurnal Penelitian dan Pengkajian Ilmu Pendidikan: E-Saintika, 4(2), 173-182.

Kemendikbud. (2017). Panduan Penilaian oleh Pendidik dan Satuan Pendidikan untuk Sekolah Menengah Atas. Jakarta: Direktorat Jenderal Pendidikan Dasar dan Menengah.

Kusuma, M. D. (2017). The Development of Higher Order Thinking Skill (HOTS) Instrument Assessment in Physics Study. Journal of Research and Method in Education, 7(1), 26-32.

Merta, I. W., Lestari, N., & Setiadi, D. (2019). Teknik Penyusunan Instrumen Higher Order Thinking Skills (HOTS) Bagi Guru-Guru SMP Rayon 7 Mataram. Jurnal Pendidikan dan Pengabdian Masyarakat, 2(1), 1-10.

OECD. (2023). PISA 2022 Assessment and Analytical Framework PISA. Paris: OECD.

Prajoko, S., Erawati, A., Amali, F. W., Zunaena, M., & Arganingtias, I. (2022). The Development of High Order Thinking Skills Test Based on Google Forms on Cell Biology Materials. Journal Of Biology Education, 5(1), 64-79.

Retnawati, H., Djidu, H., Kartianom, A., & Anazifa, R. D. (2018). Teachers’ Knowledge about Higher-Order Thinking Skills and Its Learning Strategy. Problems of Education in the 21st Century, 76(2), 215.

Sani, A. (2019). Pembelajaran Berbasis HOTS (Higher Order Thinking Skills). Tangerang: Tira Smart.

Thomas, A., & Thorne, G. (2009). How To Increase Higher Level Thinking. Metarie, LA: Center for Development and Learning, 1-17. From The Center for Literacy & Learning.

Ulfa, M., & Kuswanti, N. (2021). Development of Assessment Instrument based on Higher Order Thinking Skills of Respiratory System of Grade XI of Senior High School. Berkala Ilmiah Pendidikan Biologi (BioEdu), 10(1), 1-11.

Umami, R., Rusdi, M., & Kamid. (2021). Pengembangan Instrumen Tes Untuk Mengukur Higher Order Thinking Skills (HOTS) Berorientasi Programme for International Student Asessment (PISA) pada Peserta Didik. JP3M, 7(1), 57-68.

Widana, I. W. (2017). Modul Penyusunan Soal Higher Order Thinking Skill. Jakarta: Direktorat Pembinaan SMA, Kemendikbud.

Zainuri, A., Aquami, & Annur, S. (2021). Evaluasi Pendidikan (Kajian Teoritik). Pasuruan: Qiara Media.