JIPS 6 (2) (2025)

Jurnal Inovasi Pendidikan dan Sains

E-ISSN 2721-9119

https://ejournal.unwmataram.ac.id/index.php/JIPS

Development of Assessment Instruments for Higher Order Thinking Skills (HOTS) on Ecosystem Material

Yenny Nurmawaddah[1*], Agil Al Idrus[2], Kusmiyati[3]

[1,2,3]Biology Education Study Program, Faculty of Teacher Training and Education, Universitas Mataram, Indonesia

---

Corresponding Author:
Author Name*: Yenny Nurmawaddah
Email*: yennynurmawaddah12@gmail.com

Accepted: April 23th 2025. Approved: June 18th 2025. Published: June 25th 2025

ABSTRACT

Evaluation of learning in secondary schools is still dominated by test items that only measure memorization skills, thus failing to encourage students' Higher Order Thinking Skills (HOTS), as evidenced by research findings showing that students' achievement in the ecosystem topic remains relatively low, with 65% of 30 students failing to reach the minimum mastery criteria (KKM). This study aims to develop and evaluate the feasibility of HOTS assessment instruments based on expert validation and empirical testing. The study used a Research and Development (R&D) approach using the Plomp model, involving 115 Grade X students from two senior high schools in Labuapi District. The expert validation results indicated an instrument feasibility level of 86.39%, which falls into the very feasible category. Empirical testing showed that all test items were valid and showed high reliability (α = 0.959). The analysis of item difficulty levels varied, with a distribution of test items categorized as easy (12%), moderate (64%), and difficult (24%), while the item discrimination index showed that the majority of test items were in the very good category (64%). The results obtained indicate that the developed HOTS assessment instrument is feasible and can be used to stimulate and enhance students' Higher Order Thinking Skills on the ecosystem topic.

Keywords: development, assessment instrument, HOTS

INTRODUCTION

Higher Order Thinking Skills (HOTS) refer to the ability that emphasizes activities such as evaluating, connecting, and synthesizing various types of information [1]. Higher-order thinking involves the ability not only to understand information or lessons taught in school, but also to apply the acquired knowledge in everyday life [2]. Education has consistently emphasized teaching students to develop and enhance Higher Order Thinking Skills (HOTS); however, the assessment instruments used to measure these skills have remained underdeveloped. This condition has inevitably affected students' thinking abilities, as evidenced by a study conducted by Amaliah, which showed that student achievement in the ecosystem topic was still relatively low, with 65% of 30 students failing to reach the minimum mastery criteria (Indonesian: *Kriteria Ketuntasan Minimum*; KKM) [3].Learning activities commonly observed in classrooms have primarily involved teachers delivering instructional content through lectures, a method that emphasizes teacher activity more than student engagement. This approach leads students to become passive in developing critical or higher-order thinking, further exacerbated by the absence of final evaluation questions (HOTS) that require students to engage in higher-level thinking.

The assessment instruments used by teachers to evaluate learning outcomes are still predominantly designed to measure students' ability to memorize and recall, resulting in a limited number of test items aimed at assessing students' Higher Order Thinking Skills (HOTS). The improvement of students' learning quality can be supported through the development of HOTS-based tests using a systematic approach that includes validity testing, reliability, discrimination index, and item difficulty level, so that the learning outcomes obtained truly reflect students' Higher Order Thinking Skills (HOTS).

The implementation of learning using the *Merdeka Curriculum* develops forms of evaluation that align with educational demands, namely HOTS-based assessment [4]. The *Merdeka Curriculum* aims to produce contextual education, making learning more meaningful and not merely focused on memorizing content. This objective is in line with the development of HOTS test items, which emphasize contextual assessment so that students are not only able to recognize and understand, but also to analyze, evaluate, and create [5]. The success of HOTS test instrument development has been showed and carried out by several researchers; Maharani et al. stated that the Higher Order Thinking Skills (HOTS) assessment instrument on biology material was considered feasible

to be used as a tool for measuring students' higher order thinking abilities, as it met the requirements of content, construct, and language validity, as well as time allocation and test instructions. This was supported by the average percentage score of 85.0% in the expert assessment, which falls into the very feasible category [6]. Another study conducted by Rachma and Arsisari also showed that test items are considered valid and reliable if they meet a validity level above 0.5 and a reliability level of 0.86, with moderate and high difficulty levels, and question item discrimination categorized as good and adequate [7].

Based on these findings, the researcher identified a similar issue in schools located in Labuapi District, namely SMA Negeri 1 Labuapi and SMA Negeri 2 Labuapi. The results of observations conducted by the researcher in collaboration with biology subject teachers at each school revealed that both SMA Negeri 1 Labuapi and SMA Negeri 2 Labuapi had not fully implemented HOTS-based evaluation test items in their learning activities. Teachers prioritize test items with directives such as explain, mention, or write. In practice, teachers generally include only one or two HOTS test items in each evaluation, while most of the other items still fall into the remembering category (LOTS). This aligns with the findings of Wardani in five senior high schools in Surakarta, which showed that 90.5% of test items on the ecosystem material were still at the LOTS (Lower Order Thinking Skills) level, while only 8.4% fell into the HOTS (Higher Order Thinking Skills) category [8]. This indicates a lack of instruments that stimulate higher-order thinking.

One of the contributing factors is the low ability of students to answer such test items, leading teachers to tend to avoid the extensive use of HOTS instruments as the primary evaluation tool in learning. In fact, the implementation of HOTS test items is essential in enhancing students' critical, analytical, and creative thinking abilities as part of efforts to improve the quality of education. Therefore, the researcher will develop a Higher Order Thinking Skills (HOTS) assessment instrument on the ecosystem topic that is feasible for use based on validator evaluations and meets instrument quality standards in terms of validity, reliability, difficulty level, and item discrimination. This instrument is intended to serve as a supporting tool in learning evaluation activities that can train students to utilize their Higher Order Thinking Skills (HOTS).

RESEARCH METHOD

The type of research conducted was Research and Development using the Plomp model, which consisted of five stages, starting from: the preliminary investigation stage, the design stage, the realization/construction stage, the test, evaluation, and revision stage, and the implementation stage [9].

1. The preliminary investigation stage was carried out to identify the problems to be addressed through product development. The initial step involved conducting observations to collect several references related to the research, such as test instruments for Higher Order Thinking Skills (HOTS). After gathering the references, the next step was to examine learning tools and conduct unstructured interviews with Grade X biology subject teachers.
2. The design stage involved determining the test specifications as a prerequisite for designing an instrument, which included identifying the Learning Outcomes (Capaian Pembelajaran/CP) and Learning Objective Flow (Alur Tujuan Pembelajaran/ATP), as well as selecting appropriate material. Next, learning indicators were established based on Bloom's Taxonomy, particularly at the C4 (analyzing), C5 (evaluating), and C6 (creating) levels. After that, a test blueprint was developed to reflect these cognitive indicators, and a validation sheet was created to assess the developed instrument.
3. The realization/construction stage aimed to develop the HOTS assessment instrument on ecosystem material for the even semester following the test blueprint that had been previously prepared during the design stage.
4. The test, evaluation, and revision stage was conducted to validate the initial design of the assessment instrument by expert validators. The validation was carried out by two experts, namely a subject matter expert and an assessment instrument expert. If the validation results indicated that the instrument was valid without requiring revision, it proceeded to small group testing. If it was valid but required minor revisions, improvements were made before the trial. However, if the instrument was deemed invalid, it had to be thoroughly revised until it became feasible, and then revalidated. This process could be repeated until the instrument was declared valid.
5. The implementation stage aimed to measure the effectiveness of the HOTS assessment instrument. The HOTS test items that were declared valid and reliable in the limited-scale trial were then subjected to field testing involving 115 Grade X students from SMAN 1 Labuapi and SMAN 2 Labuapi, who served as the research subjects. The collected data were analyzed to evaluate the quality of the instrument through tests of validity, reliability, item discrimination, and difficulty level,to ensure the feasibility of the HOTS assessment instrument being used.

The sampling method was carried out using stratified random sampling, and the sample size was calculated using the Slovin formula [10].

$$n = \frac{N}{N.d^2 + 1}$$

Description:
n : number of samples
N : total population
d^2 : predetermined margin of error.

Based on the calculation results with a margin of error (d^2) of 5%, the number of research samples in the large-scale trial was 115 students from two different schools, namely SMAN 1 Labuapi and SMAN 2 Labuapi. This study involved two subject groups: a small group

trial and a field test. For the small group trial, 6 students from class X^a of SMAN 2 Labuapi were involved. The class selection was based on the results of observations and input from the subject teacher, which indicated that students in the selected class had diverse abilities and characteristics that aligned with the needs of the HOTS instrument development design.

The instruments used in this study consisted of two types, namely: (a) Expert validation instruments in the form of questionnaires administered to two validators, comprising lecturers who are experts in subject matter and in test instrument development. The purpose of this questionnaire was to assess the quality of the test instrument. The results of this validation served as a reference for revising and refining the instrument before it was administered to students. (b) Test instrument analysis involving students as subjects, which aimed to evaluate the feasibility of the test instrument. This analysis included tests of validity, reliability, item difficulty level, and item discrimination.

1. Expert Validity Analysis (Instrument Feasibility)

The data analysis technique for the validation questionnaire was conducted using a Likert scale to assess the feasibility of the developed assessment instrument. Each item in the questionnaire was accompanied by five response options: Very Good (VG) with a score of 5, Good (G) with a score of 4, Fair (F) with a score of 3, Poor (P) with a score of 2, and Very Poor (VP) with a score of 1. The total score from the validators was calculated using the following formula[11]:

$$P = \frac{f}{n} x\ 100\ \%$$

Description:
P : Percentage score from the questionnaire
f : Score obtained
n :Maximum possible score

Table1.Feasibility Criteria

Feasibility Percentage (%)	Interpretation Criteria
0–20	Very Infeasible
21–40	Infeasible
41–60	Moderately Feasible
61–80	Feasible
81–100	Very Feasible

2. Analysis of HOTS Assessment Instrument Quality

The data analysis technique for evaluating the quality of the HOTS assessment instrument, namely item validity, reliability, difficulty level, and item discrimination, was conducted with the assistance of SPSS version 20 (Statistical Program for Social Science). The data analysis technique for students' answer scores is described as follows:

a. Item Validity Test

The test item validity test was conducted using the product-moment correlation formula. The following is the formula used to calculate validity[12].

$$r_{XY} = \frac{N\ \sum XY - (\sum X)\ (\sum Y)}{\sqrt{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y^2 - (\sum Y)^2\}}}$$

Description :
r_{xy} : correlation coefficient between variable X and variable Y
N : number of respondents/test participants
$\sum X$: item score
$\sum Y$: total score
$\sum XY$: the result of multiplying the item score by the total score

If the calculated correlation coefficient ($r_{calculated}$) > the critical value (r_{table}), the item is considered valid
If the calculated correlation coefficient ($r_{calculated}$) < the critical value (r_{table}), the item is considered invalid

b. Reliability Test

Test reliability indicates the extent to which a measuring instrument can produce consistent results when used repeatedly under the same conditions. The following formula is used to calculate reliability[13]:

$$r_i = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{S_i} 2}{s_t^2}\right)$$

Description:
r_i : Cronbach's alpha reliability coefficient
k : number of test items
$\sum s_i^2$: sum of the variance of each item
s_t^2 : total variance

Table2. Interpretation of Reliability Coefficient

Reliability Coefficient	Interpretation
0.81–1.00	Very High
0.61–0.80	High
0.41–0.60	Moderate
0.21–0.40	Low
0.00–0.20	Very Low

c. Difficulty Level Test

The formula for calculating the difficulty level of a question item is as follows[14]:

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

Description :
P : Difficulty Index
B : Number of students who answered the item correctly
JS : Total number of students taking the test

Table3. Difficulty Level Criteria

Value of P	Interpretation
0.00 – 0.30	Difficult
0.31 – 0.70	Moderate
0.71 – 1.00	Easy

d. Discrimination Index Test
The formula for calculating the discrimination index is as follows[14]:

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} = P_A - P_B$$

Description:
D : Discrimination index (item discrimination score)
J_A : Number of students in the upper group
J_B : Number of students in the lower group
B_A : Number of students in the upper group who answered correctly
B_B : Number of students in the lower group who answered correctly
$P_A = \frac{B_A}{J_A}$= Proportion of testees in the upper group who answered the corresponding test item correctly
$P_B = \frac{B_B}{J_B}$ =Proportion of testees in the lower group who answered the item incorrectly

Table4. Interpretation of Discrimination Index

Discrimination Index	Category
0.00 – 0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.70	Good
0.71 – 1.00	Very Good

RESULTS AND DISCUSSION

1. Expert Validation Results of the Higher Order Thinking Skills (HOTS) Assessment Instrument

The HOTS assessment instrument, which had passed through the realization stage, was then evaluated through a validation process by experts. Each validator was given a questionnaire containing several items to assess and provide comments and suggestions on the developed assessment instrument. The validators who assessed the HOTS assessment instrument consisted of two Biology Education lecturers from Mataram University.

The feasibility of the developed HOTS assessment instrument was measured using a questionnaire provided to the validators. The questionnaire covered aspects of content, construct, and language. At the final stage of validation, each validator provided a conclusion regarding the overall feasibility of the assessment instrument. The following presents the calculated questionnaire data, showing the percentage of validation results from the experts.

Table5. Expert Validation Results

Aspect	Percentage (%)	Category
Content	87.5	Very Feasible
Construct	85.0	Very Feasible
Language	86.67	Very Feasible
Overall Evaluation	**86.39**	**Very Feasible**

The expert validation result for the content aspect, as shown in Table 5 above, reached a percentage of 87.5% and was categorized as very feasible. The analysis of the content aspect aimed to ensure the alignment between the material and the predetermined learning objectives and indicators. This is in line with the statement by Solihatul and Abidin, who emphasized that in the process of developing test items, it is essential to refer to the indicators so that the constructed items correspond to and are consistent with the aspects being measured [15]. Therefore, achieving a very feasible category in the content aspect is an acceptable outcome and aligns with the quality standards of the developed assessment instrument.

Furthermore, the expert validation result for the presentation/construct aspect obtained a percentage of 85% and was categorized as very feasible. This result was provided by the subject matter expert (Drs. Lalu Japa, M.Si., St.), who stated that the presentation/construct aspect was appropriate. Meanwhile, the language aspect received a percentage of 86.67% and was also categorized as very feasible. This result was given by the assessment instrument expert (Dr. Jamaluddin, M.Pd), who stated that the language aspect had applied proper and correct linguistic rules.

The average percentage of expert validation results for the content, construct, and language aspects was 86.39% and categorized as very feasible. This is because the developed test items covered the level of Higher Order Thinking Skills (HOTS), reflected reasoning processes, were constructed based on contextual problems, contained implicit answers, and were written using communicative language. Nevertheless, several test items needed to

be revised before proceeding to the trial stage. Suggestions provided by the validators included: improving the wording of the questions to avoid excessive length and to comply with the rules of the Indonesian Spelling System (PUEBI), as well as adjusting the answer options to follow alphabetical order.

2. Results of the Higher Order Thinking Skills (HOTS) Assessment Instrument Quality Test

a. Small Group Trial

The small group trial was conducted to minimize errors before the field test. This trial was carried out on a population that was not included in the sample, consisting of 6 students. The item validity test was conducted by comparing the r_{table} and $r_{calculated}$ values. The r_{table} value used was 0.729, determined based on a 5% significance level or a 95% confidence level.

The results of the validity test showed that all 25 multiple-choice test items were classified as valid. The reliability coefficient of the test items was 0.988, which falls under the interpretation of 'very high'. Based on the analysis of item difficulty level, the questions were not categorized as easy, difficult, or very difficult; instead, the 25 items administered to 6 students were classified as 'moderate'. These small group trial results are in line with the findings of Nurhalimah et al., who stated that the quality of a good test item is not only indicated by the fulfillment of validity and reliability aspects, but also by a moderate difficulty level, with a difficulty index between 0.30 and 0.70 [16]. Furthermore, the results of the item discrimination analysis showed that each item fell into the 'very good' category. Therefore, it can be concluded that the developed HOTS assessment instrument is feasible for use in the large-scale trial (field test).

b. Field Test

1) Empirical Validity Test

Empirical validity or item validity was conducted by administering the developed HOTS assessment instrument to 115 respondents who served as the subjects in the field test. An item is considered valid if the $r_{calculated}$ value is greater than the r_{table} value. The validity test results were obtained by comparing the r_{table} and $r_{calculated}$ values. The r_{table}value, which was 0.154, was determined based on the degrees of freedom (df = N - 2) with a 95% confidence level. The results of the validity test indicated that all 25 multiple-choice test items were classified as valid.

Setyoningtyas and Kasmui [17] explained that this occurred because the number of students who obtained high and low scores showed a balanced proportion. The items were considered valid based on the validators' assessment, as the questions were constructed properly and covered material that accurately represented the aspects intended to be measured.

Based on the results of the validity test, the $r_{calculated}$ value for each item was higher than the r_{table} value. Thus, it can be concluded that all test items were declared valid and feasible for use. A similar study was conducted by Nur and Budijastuti, which showed that the assessment instrument tested achieved a validity percentage of 86.7%, indicating that it met the criteria of being a valid instrument overall, considering that a test instrument is categorized as good if it has a minimum validity level of 70% [18]. In addition to the high results obtained in this study, there are also findings from studies with lower outcomes. A study conducted by Rizki et al. reported that the percentage of valid items in the HOTS assessment instrument validity test was 50%, while the percentage of invalid items was also 50%, out of a total of 20 test items tested [19].

The results of the validity test are affected by students' conditions when answering the questions. Students who are in a prepared state tend to find it easier to understand the content of the questions. In addition, learning interest also plays a role in affecting the quality of responses. Students with a high interest in a particular subject generally show greater attention, making it easier for them to comprehend and solve the questions. This means that the higher the students' interest in the material being tested, the greater their level of attention and ability to understand the questions will be [20].

Test items that are declared valid can be reused for future assessment activities and can also be included in a question bank. Lestari stated that items that have been validated can be utilized as a tool to develop students' Higher Order Thinking Skills (HOTS) [21].

2) Reliability Test

Reliability calculation was conducted only on the test items that were declared valid. This is because validity is a more crucial aspect, and items that do not meet the validity criteria should not be used, as they cannot measure the intended objectives. Sarkadi stated that a reliable instrument is not necessarily valid, whereas a valid instrument is guaranteed to be reliable [22]. This statement illustrates that validity influences the reliability coefficient of an assessment instrument. On the other hand, reliability cannot affect validity, as reliability depends on the validity of the

instrument itself. Therefore, an assessment instrument must meet two essential aspects, namely validity and reliability, to produce trustworthy measurement results.

The test instrument was declared reliable based on the analysis results, which showed a reliability coefficient of 0.959. This value falls into the very high category according to the reliability interpretation by Sumardi (≥ 0.81) [23]. This indicates that the developed assessment instrument has a high level of consistency, as evidenced by the uniformity in students' response patterns to each question item. A reliable test will yield the same results if administered again under the same conditions and to the same group of subjects [24].

The reliability value of an instrument is affected by several factors, both directly and indirectly. Direct factors include the time of administration, the length of the instrument, item difficulty index, score distribution, and scoring objectivity, while indirect factors involve the clarity of instructions, supervision, and environmental conditions during testing [25]. The reliability test result of the instrument in this study was 0.959, indicating a very high level of reliability. The high level of reliability was closely related to the type of questions used, namely multiple-choice questions, which have fixed answer keys and a consistent scoring system. This is in line with the opinion of Amalia and Trimulyono, who stated that reliability is influenced by the objectivity of the instrument. In this context, objectivity refers to students with diverse abilities, resulting in varied measurement outcomes [26].

The findings of this study are in line with the findings of other studies [27], [28], which reported reliability values of 0.84 and 0.818 in the development of HOTS assessment instruments, both of which fall into the very high category. However, the reliability values obtained in those studies were lower than the reliability test results obtained in the present study.This difference can be explained through the design approach and instrument refinement carried out by the researcher using the Plomp development model, which emphasizes systematic revision based on expert validation and limited-scale testing;thus, each question item was further improved before being tested on a larger scale (field test). This approach contributed to the high consistency of the measurement results.

3) Difficulty Level Test

The difficulty level of a question refers to the probability that students at a certain level of ability can answer the question correctly, expressed in the form of an index. The difficulty index is generally presented as a percentage with a value range between 0.00 and 1.00 [29]. This indicator is used to determine whether a question falls into the *easy*, *moderate*, or *difficult* category. Ideally, a test should maintain a proportional balance among these three categories, with 30% easy items, 50% moderate items, and 20% difficult items [30]. The purpose of difficulty level analysis is to ensure that the questions used can comprehensively reflect students' levels of ability [31].

The determination of difficulty level is based on the average ability of all students, not on individual abilities. To determine the difficulty level of a question, it can be calculated by measuring the percentage of students who answered correctly. The higher the percentage, the easier the question is considered to be. Conversely, if the percentage is low, the question is considered difficult [32]. The distribution of test items based on difficulty level is presented in Table 6 below:

Table6. Distribution of Test Items Based on Difficulty Level

Parameter	Category	Item Numbers	Number of Items	Percentage%
Difficulty Level	Easy (0.71–1.00)	1,14,21	3	12%
	Moderate (0.31–0.70)	2,3,4,5,6,7,9,11,13,15,16,18,19,22,23, 24	16	64%
	Difficult (0.00–0.30)	8,10,12, 17,20,25	6	24%

The difficulty level of the questions was determined based on the calculation results obtained using SPSS version 20. According to the data presented above, out of 25 test items, those categorized as easy accounted for 12% or 3 items. Meanwhile, the moderate category consisted of 16 items,or 64%, and the difficult category included 6 items, representing 24%. Overall, the difficulty level of the multiple-choice questions fell into the moderate category. This finding is consistent with the results of a study by Pradita et al., which showed that based on the difficulty scale, the majority of

items were in the moderate category, with a percentage of 85% or 34 out of a total of 40 test items [33]. Furthermore, a study conducted by Septiani et al. found that 80% (24 items) out of a total of 30 questions had a difficulty index in the range of 0.31–0.70 and were classified within the moderate difficulty level criteria [34].

In addition to the findings indicating a high number of items in the moderate category, a study by Putri et al. reported a lower number of items in the moderate category compared to the present study, with 21 items classified as difficult, 8 items as moderate, and 1 item as easy [35]. The difference between these findings and the results of the present study can be explained by variations in the methods applied, the scope of the material, respondent characteristics, and other contributing factors.

Oktaviana [36] explained that a good question item has a moderate level of difficulty, meaning it is neither too easy nor too difficult. The items found in this study are considered ideal because they are neither overly easy nor overly difficult. Meanwhile, items that fall into the too-easy or too-difficult categories indicate that, in terms of content, the questions do not fully represent the material that has been taught. Questions that are too easy will not stimulate students' thinking abilities, while questions that are too difficult may cause students to lose motivation to try, as they exceed the students' level of ability.

Referring to the ideal criteria, the distribution of the difficulty levels of the developed questions has shown a pattern consistent with the standard. The analysis results indicated that there were 3 items (12%) in the easy category, 16 items (64%) in the moderate category, and 6 items (24%) in the difficult category. This distribution aligns with the ideal proportion proposed by Sudjana, as cited by Warju et al., which follows a 3-5-2 ratio, or 30% of items in the easy category, 50% in the moderate category, and 20% in the difficult category [37].

4) Discrimination Index Test

Items with a discrimination index value approaching 0.00 indicate that the question has low discrimination power. Conversely, if the discrimination index value approaches 1.00, the item is categorized as having very good discrimination power [38]. A question item is considered to be of good quality if it has a minimum discrimination index value of 0.20, which consideredas the fair category. The distribution of test items based on their discrimination index is presented in Table 9 below:

Table7. Distribution of Test Items Based on Discrimination Index

Parameter	Category	Item Numbers	Number of Items	Percentage (%)
Discrimination Index	0.71–1.00 (very good)	2,3,4,5,6,7,9,11,13,15,16,18,19,22,23,24	16	64
	0.41–0.70 (good)	8,10,12,14,17,20,25	7	28
	0.20–0.40 (fair)	1,21	2	8
	0.00–0.19 (poor)	-	-	0

Based on the summary of the data above, it can be seen that 16 out of 25 test items (64%) had very good discrimination power, 7 out of 25 items (28%) were in the good category, while 2 items (8%) were in the fair category. The analysis results showed that there were no items categorized as having poor discrimination power (0.00–0.20). This finding is consistent with the results of a study by Mirza et al., which showed that 4 items (26.67%) were in the fair category, 5 items (33.33%) were in the good category, and 6 items (40.00%) were in the very good category, with no items falling into the poor discrimination category [39]. Based on the research findings by Ayu et al., a question item is considered to meet the feasibility criteria if its discrimination index falls within the fair, good, or very good categories [40].

Therefore, it can be concluded that this HOTS assessment instrument has a reasonably good quality in distinguishing students' levels of ability, as the majority of the test items fall into the good and very good categories. Although some items remain in the fair category, revisions can be made to improve their discrimination power and enhance the overall quality of the test.

CONCLUSION

Based on the results of the research and the discussion conducted by the researcher, it can be concluded that the development of the Higher Order

Thinking Skills (HOTS) assessment instrument on the ecosystem material for Grade X senior high school students in Labuapi District is very feasible, with a feasibility percentage of 86.39%. In addition, the validity test results showed that 25 test items met the criteria for validity. The reliability test indicated that the instrument had very high reliability, with a reliability coefficient of 0.959. The item difficulty analysis revealed that 3 items (12%) were in the easy category, 16 items (64%) in the moderate category, and 6 items (24%) in the difficult category. Meanwhile, based on the discrimination index analysis, 16 items (64%) were categorized as very good, 7 items (28%) as good, and 2 items (8%) as fair.

ACKNOWLEDGEMENTS

The author expresses gratitude to Allah SWT for the ease and smooth progress granted in completing this research. The author also extends heartfelt thanks to both parents and family for their unwavering prayers, support, and encouragement. Furthermore, the highest appreciation is addressed to the academic advisor for patiently providing guidance, direction, assistance, and valuable input throughout the process of composing and completing this research. Sincere thanks are also conveyed to the principals, biology teachers, staff, and the entire academic community of SMAN 1 Labuapi and SMAN 2 Labuapi for their support and cooperation during the implementation of this study. Finally, the author wishes to thank fellow students Nasipatul Pajriati, Yeni Susanti, and Aulia Zuhrianti for their assistance in conducting the research.

REFERENCES

[1] S. Yuliana, T. Sunanti, and Kintoko, "Analisis Higher Order Thinking Skill (HOTS) Siswa Dalam Menyelesaikan Soal Cerita Matematika," *RIEMANN: Research of Mathematics and Mathematics Education*, vol. 6, no. 1, pp. 118-126, 2024.

[2] N. Zebua, "Studi Literatur: Peranan Higher Order Thinking Skills Dalam Proses Pembelajaran," *Jurnal Inovasi Pendidikan*, vol. 1, no. 2, pp. 92-100, 2024. doi: https://doi.org/10.62383/edukasi.v1i2.110.

[3] N. Amaliah, "Pengaruh Model Pembelajaran Search, Solve, Create, and Share (SSCS) Terhadap HOTS (High Order Thinking Skills) Biologi Siswa Kelas X Pada Materi Ekosistem Di SMA Negeri 3 Gowa". Skripsi, Univ. Muhammadiyah Makassar, 2021.

[4] N. Afdilani Sitompul, N. Anas and L. Nur Kamalia Siegar, "Pengembangan LKPD Berbasis Discovery Learning Pada Materi Ekosistem Untuk Meningkatkan HOTS Siswa Kelas X SMA," *Spiaetus: Jurnal Biologi dan Pendidikan Biologi,* pp. 243-260, 2023.

[5] D. Wadini, "Pengembangan Soal Higher Order Thinking Skills (HOTS) Berbasis Literasi Numerasi Sains Biologi Pada Materi Ekosistem Kelas X SMA," Skripsi, Univ. Sriwijaya, 2025.

[6] R. Maharani Nasution, Hasanuddin and F. Harahap, "Pengembangan Instrumen Penilaian Hasil Belajar Berbasis HOTS Pada Materi Biologi Semester Ganjil Kelas XI SMA," *Jurnal Kependidikan*, vol. 13, no. 3, pp. 3513-3521, 2024.

[7] E. Rachma Kurniasi and A. Arsisari, "Pengembangan Instrumen Pengukur Higher Order Thinking Skills (HOTS) Matematika Pada Siswa Sekolah Menengah Pertama," *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, vol. 9, no. 4, pp. 1213-1222, 2021.

[8] K. Wardani, "Kelayakan Instrumen Pengembangan Penilaian Higher Order Thinking Skills Siswa SMA Pada Materi Ekosistem," *Jurnal Pendidikan Sains (JPS)*, vol. 6, no. 2, pp. 21-31, 2018.

[9] Fadillah, "Pengembangan InstrumenPenilaian Untuk Mengukur Keterampilan Proses Sains Siswa SMA," *Didaktika Biologi: Jurnal Penelitian Pendidikan Biologi,*vol. 1, no. 2, pp. 123-134, 2017.

[10] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: CV. Alfabeta, 2017.

[11] S. Arikunto, *Dasar-Dasar Evaluasi Pendidikan Edisi 3*. Jakarta: Bumi Aksara, 2018.

[12] R. Ananda and M. Fadhli, *StatistikaPendidikan*.Medan: CV. Widya Puspita, 2018.

[13] U. Sekaran, *Metode Penelitian Untuk Bisnis*. Jakarta: Salemba Empat, 2017.

[14] L. Umi Fatimah and K. Alfath, "Analisis Kesukaran Soal, Daya Pembeda dan Fungsi Distraktor," *JurnalKomunikasi dan Pendidikan Islam*, vol. 8, no. 2, pp. 37-64, 2019.

[15] A. Solihatul and Z. Abidin," Analisis Kesesuaian Butir Soal Ulangan Harian dengan Indikator Serta Evaluasi Pelaksanaan Penilaian Autentik Pada Materi Pecahan Kelas 4 Di MI Sudirman," Prosiding SNDIK Magister Pendidikan Dasar UMS, pp.142-151, 2020.

[16] S. Nurhalimah, Y. Hidayati, I. Rosidi and W. Puspita Hadi, "Hubungan Antara Validitas Item Dengan Daya Pembeda dan Tingkat Kesukaran Soal Pilihan Ganda PAS," *Jurnal Natural Science Educational Research,* vol. 4, no. 3, pp. 1-9, 2022.

[17] R. Setyoningtyas, and Kasmui, "Pengembangan Quizizz-Assissted Tesr Berbasis Literasi Sains Pada Materi Larutan Elektrolit Nonelektrolit," *Journal of Chemistry In Education*, vol. 9, no. 2, pp. 1-7, 2020.

[18] Z. Nur Azizah and W. Budijastuti, "Pengembangan Instrumen Penilaian Untuk Mengukur Keterampilan Literasi Sains Pada Submateri Sistem Peredaran Darah Manusia," *Jurnal BioEdu*, vol. 11, no. 1, pp. 89-97, 2022.

[19] D. Rizki, F. Abiyyu Pakarti Almay, and S. Dwijayanti Ramadani, "Pengembangan Instrumen Penilaian Higher Order Thinking Skills (HOTS) Pada Materi Ekologi SMA," *Biodik: Jurnal Ilmiah Pendidikan Biologi*, vol. 10, no. 4, pp. 691-702, 2024, doi: https://doi.org/10.22437/biodik.v10i4.36741.

[20] A. Fajar Rini and W. Budijastuti, "Pengembangan Instrumen Soal *HOTS* Untuk Mengukur Keterampilan Pemecahan Masalah Pada Materi Sistem Gerak Manusia," *Jurnal BioEdu*, vol. 11, no. 1, pp. 127-137, 2022.

[21] W. Lestari, "Pengembangan Instrumen Multiple Choice Terbuka Berbasis HOTS dengan Pendekatan Literasi Sains Untuk Mengukur Kemampuan Berpikir Tingkat Tinggi Siswa Kelas X SMAN Karangpandan Pada Materi Gerak Harmonik," Skripsi, Univ. Islam Negeri Walisongo, 2019.
[22] Sarkadi, *Tahapan Penelitian Pembelajaran Berdasarkan Kurikulum 2013*. Surabaya: CV. Jakad Media Publishing, 2020.
[23] Sumardi, *Teknik Pengukuran dan Hasil Belajar*. Yogyakarta: DEEPUBLISH, 2020.
[24] M. Hisyam Baidlowi, Sunarmi and Sulistijono, "Pengembangan Instrumen Soal *Essay* Tipe *Higher Order Thibking Skills* (HOTS) Materi Struktur Jaringan dan Fungsi Organ Pada Tumbuhan Kelas XI SMAN 1 Tumpang," *Jurnal Pendidikan Biologi*, vol. 10, no. 2, pp. 57-65, 2019.
[25] Z. Suwitaningsih and S. Indana, "Pengembangan Instrumen Penilaian Akhir Semester (PAS) Mata Pelajaran Biologi Pada Kelas X Di MAN Sidoarjo," *Jurnal Ilmiah Pnedidikan Biologi*, vol. 7, no. 2, pp. 298-303, 2018.
[26] P. Amalia Salsabila and G. Trimulyono, "Pengembangan Instrumen Penilaian HOTS Materi Virus Untuk Mengukur Kemampuan Berpikir Kritis Siswa Kelas X SMA," *Jurnal BioEdu*, vol. 12, no. 2, pp. 287-297, 2023.
[27] Mahrawi, Usman and M. Nur Avianti, "Pengembangan Instrumen Asesmen Critical Thinking Skills pada Materi Sistem Ekskresi," *Indonesian Journal of Mathematics and Natural* Science, vol. 2, no. 2, pp. 80-94, 2021, doi: http://doi.org/10.36719/mass.v2i2.72
[28] F. Fidia, R. Pratiwi Puspitawati and P. Yakub, "Pengembangan Instrumen Soal Higher Order Thinking Skills (HOTS) Materi Jaringan dan Organ Pada Tumbuhan Kelas XI SMA," *Jurnal BioEdu*, vol. 11, no. 3, pp. 745-754, 2022.
[29] Hendrayadi, M. Kustati and R. Amelia, "Analisis Ulangan Harian Mata Pelajaran PAI di SMA Negeri 10 Padang Tahun Pelajaran 2023/2024 (Telaah Terhadap Reliabilitas, Daya Beda dan Tingkat Kesukaran Menggunakan Software Anates)," *Jurnal Review Pendidikan dan Pengajaran*, vol. 7, no. 3, pp. 6954-6961, 2024.
[30] N. Zalfa Zuhri and S. Tatang, "Analisis Validitas, Reliabilitas, dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis *Artificial Intelligence* (AI) melalui Platform Question Well," *Jurnal Pendidikan dan Pembelajaran Indonesia*, vol. 4, no. 2, pp. 693-704, 2024.
[31] W. Ramadhan, F. Malahati, K. Romadhon and S. Ramadhan, "Analisis Butir Soal Tipe Multiple Choice Question pada Penilaian Harian Sekolah Dasar," *Jurnal Penelitian Pendidikan dan Pembelajaran*, vol. 10, no. 2, pp. 93-105, 2023.
[32] N. Permatasari and S. Indana, "Pengembangan Tes Elektronik *(E-Test)* Materi Perubahan Lingkungan Untuk Mengukur Kemampuan *Probelem Solving* Siswa Kelas X SMA," *Jurnal BioEdu*, vol. 9, no.1, pp. 319-324, 2020.
[33] E. Pradita, P. Megawanti and Yulianingsih, "Analisis Tingkat Kesukaran, Daya Pembeda, dan Fungsi Distraktor PTS Matematika SMPN Jakarata," *Jurnal Ilmiah Mahasiswa Pendidikan Matematika*, vol. 3, no. 1, pp. 109-118, 2023.
[34] D. Septiani, Y. Widiyawati and I. Nurwalidah, "Pengembangan Instrumen Tes Ringanrasi Sains PISA Aspek Menjelaskan Fenomena Ilmiah Kelas VII," *Jurnal Pendidikan dan Aplikasi Sains*, vol. 1, no. 2, pp. 46-55, 2019.
[35] I. Putri Agustina, F. Suzanti and M. Natalina L, "Pengembangan Instrumen Tes Berbasis Literasi Sains Pada Materi Ekosistem Kelas X SMA/MA," *Jurnal Biogenesis*, vol. 19, no. 1, pp. 17-32, 2023.
[36] V. Oktaviana Bano, D. Ndamung Marambaawang and Y. Njoeroemana, "Analisis Kriteria Butir Soal Ujian Sekolah Mata Pelajaran IPA di SMP Negeri 1 Waingapu," *Ideas: Jurnal Pendidikan, Sosial, dan* Budaya. vol. 8, no. 1, pp. 145-152, 2022.
[37] Warju, S. Rizki Ariyanto, Soeryanto and R. Ado Trisna. "Analisis Kualitas Butir Soal Tipe HOTS Pada Kompetensi Sistem REM Siswa Di Sekolah Menengah Kejuruan," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 17, no.1, pp. 95-104, 2020.
[38] S. P. Radja, V. Oktaviana Bano and A. Tamu Ina, "Analisis Kualitas Butir Soal Hasil Belajar Siswa Berdasarkan Tingkat Kesukaran, Daya Beda dan Efektivitas Pengecoh di SMAN 1 Panawai," *Jurnal Edusavana*, vol. 1, no. 1, pp. 30-41, 2023.
[39] N. Mirza Sabrina, K. Nayla Dewinta Hemi, S. Hildayati and L. Hakim, "Penerapan Aplikasi Anates Dalam Analisis Soal Pilihan Ganda di SMKN 2 Tuban," *Jurnal Pendidikan Ilmiah Transformatif*, vol. 8, no. 12, pp. 68-76, 2024.
[40] W. Ayu Fietri, Lutfi, Syamzurizal and Zulyusri, "Analisis Butir Soal Biologi Kelas VIII Madrasah Tsanawiyah Negeri 6 Kerinci," *Jurnal Pendidikan Biologi Undiksha*, vol. 8, no. 2, pp. 50-60, 2021.