Tadris: Jurnal Keguruan dan Ilmu Tarbiyah 7 (1): 47-61 (2022)
DOI: 10.24042/tadris.v7i1.10939

TADRIS: JURNAL KEGURUAN DAN ILMU TARBIYAH
(Tadris: Journal of Education and Teacher Training)
P-ISSN: 2301-7562 | E-ISSN: 2579-7964
ejournal.radenintan.ac.id/index.php/tadris/index

The Development of Testlet Assessment Instrument Model Integrated with E-ujian Website to Measure the Higher-Order Thinking Skills

Kartika Dyah Permata Mega Dwijayanti*, Erna Noor Savitri
Department of Science Education, Universitas Negeri Semarang, Indonesia

Article History:
Received: December 30th, 2021
Revised: February 10th, 2022
Accepted: May 28th, 2022
Published: June 29th, 2022

Keywords:
Assessment,
E-ujian website,
HOTS,
Testlet

***Correspondence Address:**
kartikadyahpermata0@gmail.com

Abstract: This research & development research aimed to determine the feasibility and characteristics of the Testlet model assessment instrument integrated with the E-ujian website in measuring the Higher-order Thinking Skills (HOTS) of junior high school students. The researchers employed the 4D development model (define, design, develop and disseminate). However, the researchers conduct the development up to the develop stage. The research was conducted at a junior high school with 42 students as respondents. The researchers collected the data through test techniques, questionnaires, interviews, observation, and documentation. This development research utilized the descriptive-quantitative data analysis method through characteristic tests, teacher and students' questionnaire responses analysis, and analysis of students' HOTS profiles. The students' HOTS profiles based on the assessment instrument were divided into 53 percent of students in the poor HOTS category, 31 percent of students in the low HOTS category, and 16 percent of students in the high HOTS category. Concluded that the assessment instrument developed was very feasible, with a percentage of 94 percent. The assessment instrument developed had the characteristics of valid, reliable, discriminatory, disproportionate, relevant, representative, impractical, and specific. The Testlet model assessment instrument development can be used as an alternative for the distance learning assessment during the Covid-19 pandemic.

INTRODUCTION

Competency learning as a part of the 2013 curriculum contains authentic learning and assessment processes so that students can achieve attitude, knowledge, and skill competencies (Asrizal et al., 2018; Nurtanto et al., 2021). Learning objectives will be achieved if the learning process and assessment show that all learning competencies are achieved and mastered by students. Masitoh & Aedi, (2020) explain that the knowledge competence (cognitive domain), especially in the content standards of the 2013 curriculum, is designed so that students can have high-order thinking skills (HOTS). Therefore, the assessment must measure students' HOTS appropriately.

The low level of higher order thinking skills (HOTS) can be developed through the learning process and learning assessment (Istiyono et al., 2018; Widana, 2018). The standard for assessing student learning outcomes must focus on HOTS. HOTS mastery enables students to develop themselves in deciding, assessing, and solving problems appropriately. Prayitno & Nofiana, (2020) explain that there are five skills that

makeup HOTS: (1) critical thinking skills, (2) creative thinking skills, (3) problem-solving skills, (4) decision-making skills, and (5) argumentative skills.

Higher-order thinking skills (HOTS) can make students connect, manipulate, and transform their knowledge and experience to think critically and creatively (Arif & Yuhdi, 2020; Arnellis et al., 2020; Halim et al., 2018; Nisa et al., 2018). One of the benefits of using HOTS in learning is that students' information or knowledge will be stored longer than just using Lower-Order Thinking Skills (LOTS). Besides, HOTS can also make students develop themselves in deciding and solving problems appropriately. The 2015 TIMSS (Trends in International Mathematics and Science Study) showed that Indonesia is ranked 44th out of 49 countries globally. This result shows that the students' HOTS level was poor (Syamsul Hadi & Novaliyosi, 2019). Students often work on questions that tend to LOTS. Low HOTS can be caused by the model or learning method that cannot develop students' HOTS (Retnawati et al., 2018; Tan & Halili, 2015). Also, according to (Rahayuningsih & Jayanti, 2019), the low HOTS can be caused by students who are less accustomed to working on HOTS-based questions.

Assessment instruments are essential in learning. The quality of the assessment instrument directly affects the accuracy of the student's competency achievement status (Dharmawati et al., 2016; Qi & Mitchell, 2012). It can be concluded that an appropriate assessment instrument is needed to measure the HOTS of junior high school students appropriately.

One of the assessment instruments that can be used to measure students' HOTS is a Testlet. A Testlet is a collection of items containing information on the same topic where the items have been considered a single unit that shares the same problem contest (Nova et al., 2016). Lutviana et al., (2019) explain that the Testlet model questions have more practicality than the description form test because the scoring can be done objectively and is polytomic. Testlet is an assessment instrument that can combine the advantages of multiple-choices questions and descriptions (Safitri, 2020; Shiell & Slepkov, 2015; Slepkov, 2013; Wahyuni et al., 2015).

The items in the Testlet model test consist of main questions and supporting questions. The questions are made to provide information to other supporting questions. The supporting items are made to have a level of completion of the main problem so that it is expected to help teachers diagnose students' learning difficulties (Muna et al., 2017). Beside diagnosing learning difficulties, Testlet can also measure students' skills deeper (Fahmina et al., 2019). (Shiell & Slepkov, 2015; Slepkov & Shiell, 2014) explain that the correct answer to each question can give students an idea of the whole, partly or not, regarding mastery of the following questions. The scoring method used in the Testlet is the Graded Response Model (GRM).

Fahmina et al., (2019) utilize a computerized Testlet to measure the literacy skills of high school students. Computerized assessment can also be done using the help of websites. One of the websites that provide services to carry out online assessments or assessments is E-ujian. E-ujian facilitates teachers to carry out online assessments. It can be accessed via a computer or smartphone to assess anywhere. The tracking feature through the camera and sharing sessions in the exam implementation can help teachers monitor and minimize cheating committed by students during examinations.

The development of Testlet model assessment instruments has been widely carried out, one of which is the development carried out by (Nova et al., 2016). They have succeeded in

developing an assessment instrument on temperature and heat material for the tenth-grade high school students. Other research has also been carried out by (Muna et al., 2017), who succeeded in developing a Testlet model test instrument to detect students' learning difficulties on buffer solutions topic. (Damayanti, 2017) developed a Testlet model assessment instrument to measure higher-order thinking skills of high school students on electrochemical material. The higher-order thinking skills indicators were the skills to analyze, evaluate, create, think critically, and think logically.

Based on the description, research gaps have not been investigated further. One of the gaps is further research on the Testlet model assessment instrument to measure the students' HOTS with different indicators. Therefore, the solution to this problem is to develop an integrated Testlet model assessment instrument on the E-ujian website to measure the higher-order thinking skills of junior high school students.

METHOD

This research is development research with a 4D development model. Thiagarajan's 4D development model consists of four stages: define, design, develop, and disseminate (Thiagarajan et al., 1974; Utomo & Kustijono, 2015). However, the researchers conducted the research only up to the develop stage. This development research involved lecturers and science teachers who were tasked to validate the developed product and the research subject (the seventh-grade junior high school students). The number of students who became the research subject consisted of 42 students.

Two tests were carried out at the develop stage on the developed products. First is the small-scale trial to determine the students' and teachers' responses to the Testlet model assessment instrument. There were ten students and four teachers involved in the small-scale trial. Second, a large-scale trial was conducted on 32 students to determine the characteristics of the assessment instrument developed. At the same time, the large-scale trial data was also used to analyze the students' HOTS profiles.

Instruments used to collect data in this study consisted of expert validation sheets, teacher response questionnaires, student response questionnaires, and question scripts. The experts involved were material and assessment experts. Expert validation sheets were used to assess product feasibility. Students' and teachers' questionnaire responses were used to determining the product's characteristics. The characteristics were relevance, representativeness, practicality, and specificity. On the other hand, the question text was used to determine the product's general characteristics (valid, reliability, discriminatory, and proportional) as an assessment instrument. The Testlet model was developed based on five HOTS aspects. The HOTS indicators can be seen in Table 1.

Table 1. Indicators for the Testlet Assessment Model

No.	HOTS Aspects	Indicators	Items	Total
1	Critical thinking skills (Ennis, 2011)	1. Compiling information	1.2, 7.1, 6.1, 10.1, 11.1, and 14.1	6
		2. Creating and determining self-assessment	3.2 and 14.3	2
2	Creative thinking skills (Silver, 1997)	1. Originality	11.3	1
		2. Flexibility	5.2	2
		3. Fluency	3.1, 7.2, and 12.1	3
3	Problem-solving skills (Polya, 1993)	1. Problem understanding	7.3 and 12.3	2
		2. Planning the problem-solving	1.3, 5.3, and 6.3	3
		3. Implementing the problem-	10.3	1

No.	HOTS Aspects	Indicators	Items	Total
		solving plan		
		4. Rechecking the results	12.2	1
4	Decision-making skills	1. Problem understanding	7.3 and 12.3	2
	(Wang & Ruhe, 2007)	2. decicion making	3.3 and 5.1	2
5	Argumentative skills	1. Claim	6.2 and 11.2	2
	(Toulmin & Stephen,	2. Data	1.1 and 14.2	3
	1979)	3. Rebutal	10.2	

The data used to analyze the HOTS profile of students was obtained from the results of the analysis of questions with cognitive levels C4-C6 containing indicators of the HOTS aspect. The questions developed in the Testlet model assessment instrument consisted of 15 main questions, where each main question was integrated with three supporting questions. The questions developed were questions with cognitive levels C1-C6. The questions with the HOTS cognitive level were contained in the main questions numbered 1, 3, 5, 6, 7, 10, 11, 12, and 14. Each main question was integrated with three supporting items so that there were 27 total questions used for students' HOTS analysis.

The feasibility analysis of the instrument was carried out through the percentage descriptive method. The feasibility value was obtained from the expert validation questionnaire sheet. The feasibility value obtained was then interpreted into four criteria, namely (1) highly feasible if the percentage is 81.25 < x 100.00, (2) feasible if the percentage is 62.50 < x 81.25, and (3) not feasible if the percentage is 43.75 < x 62.25, and (4) highly not feasible if the percentage is 25.00 x 43.75. The assessment instrument for the integrated Testlet model on the E-ujian website can be feasible if it has a percentage higher than 62.50 in the feasible and highly feasible criteria.

The validation was carried out by five material experts and five assessment experts. The results of expert validation were analyzed using the Aiken formula to obtain the content validity values of the items from the assessment instrument. Each item is said to be valid if the validity test results show that Vcount is greater than Vtable with a significant level of 5 % (0.78 for assessment and material experts).

The reliability of the developed Testlet model assessment instrument was analyzed using the Cronbach Alpha method. In this method, the value of the reliability coefficient alpha (r_{11}) is compared with the value of the r_{table} with a 5 % significance level. If the r_{count} is higher than r_{table}, the question is declared reliable.

The characteristics of discrimination can be analyzed by performing a discriminatory test. The discriminatory power is the skills of a question to distinguish between high-skills students and low-skills students (Arikunto, 2013). The discriminatory power test used was a discriminatory power test to assess question descriptions. Discriminatory power was obtained by comparing the difference between the upper and lower averages with the maximum score of the Tetslet model assessment instrument being tested.

The researchers interpreted the discriminatory power criteria from the results of the calculation of the discriminatory power of the items. There were four criteria used in the discriminatory power test. The question is accepted if the discriminatory power test is in the interval of $0.40 \leq D \leq 1.00$. The question is accepted but needs to be corrected if the discriminatory power test is in the interval of $0.30 \leq D \leq 0.40$, while the question needs to be corrected if it is at the interval of $0.20 \leq D \leq 0.40$. The question cannot be used if the discriminatory value is $0.00 \leq D \leq 0.20$.

Proportional analysis was performed by conducting a test of item difficulty. Rusilowati et al. (2016) state that the difficulty level test can be done by comparing the average score with the maximum score obtained on each item. The results of the difficulty level were interpreted against the existing criteria. There were three criteria for the level of difficulty: (1) Difficult if the coefficient of difficulty level is in the interval of $0.00 \leq ID \leq 0.30$, (2) Medium if the coefficient of difficulty level is in the interval of $0.30 \leq ID \leq 0.70$, and (3) Easy if the coefficient of difficulty level is in the interval of $0.70 \leq ID \leq 1.00$. Items that are good to use are those with the criteria of the interval of $0.70 \leq ID \leq 1.00$ or having moderate criteria.

Questionnaire analysis of student and teacher responses was carried out to analyze the relevant, representative, practical characteristics and specifications of the assessment instrument that had been developed. The answer to each question or statement is calculated as the Statement of Approval Level (SAL) using the representative descriptive method, comparing the total scores obtained with the total number of ideal scores.

The researchers interpreted the questionnaire assessment criteria from the results of the calculation of the level of approval of the statement. There are four criteria used, namely (1) excellent when the SAL is in the interval of $75.00 < SAL \leq 100.00$, (2) high when the SAL is in the interval of $50.00 < SAL \leq 75.00$, and (3) low when it is in the interval of $25.00 < SAL \leq 50.00$, and (4) poor when the SAL is in the interval of $0.00 \leq SAL \leq 25.00$.

HOTS analysis was carried out using the GRM method. The Graded Response Model (GRM) is a scoring model that applies a grading system to assess the answers to questions (Mateucci & Stracqualursi, 2006; Morse et al., 2012). GRM is one of the models developed to handle scoring on polytomic items (LaHuis et al., 2011; Momani, 2017). Based on the GRM scoring method, students will get a maximum score of 3 if they can complete stages 1, 2, and 3 correctly. On the other hand, they will get a score of 0 if they cannot complete the first stage correctly. The results of the large-scale trial were used to determine the students' HOTS for each indicator. The percentage of students' HOTS was then calculated using the following mathematical equation.

$$\%\ HOTS = \frac{x}{n} \times 100\%$$

Description:
x = Number of students who answered each question correctly
n = Total number of students

The percentage was then categorized based on predetermined criteria. The criteria used in determining the students' HOTS result from the adaptation of the HOTS criteria (Azmi et al., 2021). The HOTS criteria are presented in Table 2.

Table 2. The HOTS Level of Students

Percentage	Criteria
$76 \leq X \leq 100$	Excellent
$51 \leq X \leq 75$	High
$26 \leq X \leq 50$	Low
$1 \leq X \leq 25$	Poor

RESULT AND DISCUSSION

The Feasibility of Testlet Assessment Instruments Model Integrated with E-ujian Website

The feasibility of the product is determined by calculating the total score obtained from the assessment and media experts' validation using the percentage descriptive technique. The feasibility percentage from the assessment experts was 94 %, while the feasibility percentage from the material experts was 95 %, both in the highest feasible category. The experts stated that the Testlet model assessment instrument integrated with the E-ujian website to measure the higher-order thinking skills of junior high school students was highly feasible to be used in

the learning process with an average percentage of 94.5 %.

The aspects assessed by the assessment expert were the content and construct aspects. The percentage obtained for both aspects was 95%, which was declared highly feasible. Each aspect was assessed through indicators. In the content aspect, the indicator that got the highest percentage was "The relationship between competencies, Competency Achievement Indicators (GPA), questions, questions, and Bloom's Taxonomy level," with a percentage of 100%. The indicator that got the lowest percentage was the " The suitability for measuring HOTS" with a percentage of 90%.

The construct aspects of the assessment expert validation were assessed through four indicators: (1) ease of understanding, (2) systematic, (3) practicality, and (4) use of proportional typeface and font size. The indicator that received the highest percentage was "easy to understand" with a percentage of 100 %. The other three indicators get a percentage of 95% with highly feasible criteria. The assessment experts reviewed four aspects: competence, material quality, display, and language. The average percentage of all aspects assessed was 94%, with highly feasible criteria. The percentage of display indicators was 100% in the highest feasible category. Meanwhile, the lowest percentage (90%) was found in the indicator of the relationship between competencies, Competency Achievement Indicators (GPA), questions, and questions and material indicators. The language indicators contained the language assessment aspect.

Besides producing feasibility data from the assessment instrument at the validation stage, the Testlet model integrated with the E-ujian website also obtained suggestions and input. The suggestions and input were used for improvement. Improvements were made to improve the quality in terms of assessment and material. One of the improvements based on assessment experts' input related to the question grid. Improvements were made by adding a cognitive dimension mapping column. Improvements to the grid were also carried out by removing the HOTS indicator on questions with LOTS and MOTS cognitive levels. The HOTS indicator on the question was replaced with the Competency Achievement Indicator (GPA). Other improvements were also made based on the advice of the material experts. The design improvements were also carried out by improving the writing and layout of images, discourse on the main questions, supporting questions, and answer choices. Improvements to the questions on the E-ujian website page can be seen in Figure 1.

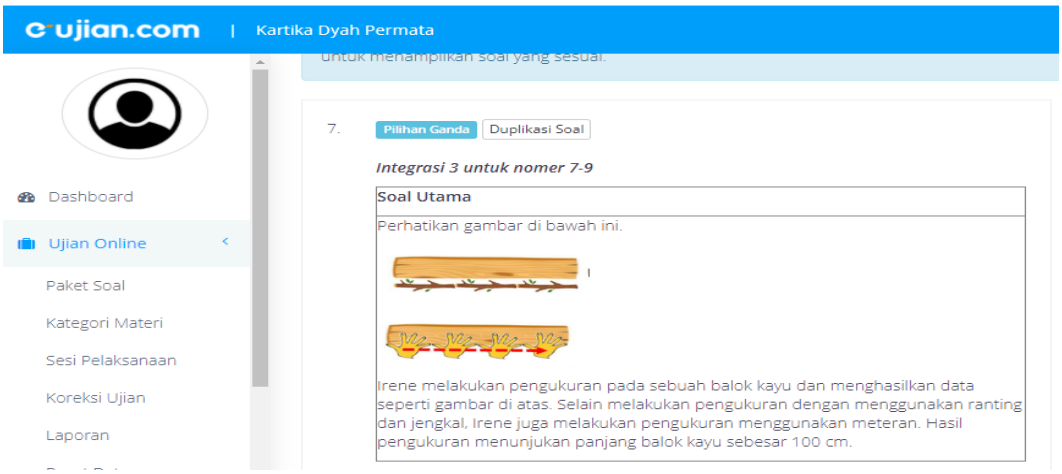

Figure 1. Improvement on the E-ujian Website

Item number three in Figure 1 was improved by increasing the size of the image to be easily observed by students. These improvements were made based on the advice of assessment experts and materials experts.

Characteristics of Assessment Instruments Testlet Model Integrated E-ujian Website

Valid

The results of the content validity analysis stated that the developed assessment instrument had high validity. (Arikunto, 2013; Hayashi et al., 2019) Explain that a test can be enriched as valid if it can measure what it wants to measure. All of the developed questions (45 items) can be declared valid. Five assessors determined the item validity with a significant level of 5%. The results showed that V_{count} was higher than V_{table} (0.87).

Reliability

The researchers analyzed the reliability using the Cronbach Alpha formula assisted by the Microsoft Excel 2016 application. Based on the results of the analysis, the reliability value (r_{11}) was 0.70 for 32 students in the large-scale trial (class VII E). (Rusilowati et al., 2016) state that the reliability of the instrument is high if the reliability coefficient (r_{11}) is 0.6 (r lower than 0.8). Based on the analysis results, the Testlet model assessment instrument can be trusted and provides the same results if used on different occasions, such as on different subjects, conditions and places, and times.

Discriminatory

Discriminatory characteristics can be known through the analysis of discriminatory tests. The data used to analyze discriminatory power was obtained at the small-scale trial stage. The average value of discriminating power for all items developed was 0.40, which was included in the high category. The results of the analysis of the discriminatory test showed that from the 15 main questions developed. Three main questions had sufficient criteria. These items are the main questions number 1, 7, and 10.

Proportional

According to (Arifin, 2012; Boopathiraj & Chellamani, 2013; Suek, 2021) the proportion of good difficulty levels is 25 % difficult questions, 50 % moderate questions, and 25 % easy questions (1:2:1 proportional comparison). The results of the proportional comparison of the developed assessment instruments were 5:4:0. Therefore, the Testlet model assessment instruments developed were declared disproportionate.

The difficulty level for all the main items developed was 0.35 in the medium criteria. The questions used to measure students' HOTS profiles had an average difficulty level of 0.29 in the difficult criteria. HOTS questions are found in the main question numbers 1, 3, 5, 6, 7, 10, 11, 12, and 14. The difficult question criteria were obtained because students were unfamiliar with Testlet-type questions and the limited learning time during distance learning, so the material provided was not studied thoroughly. (Lichtenstein & Fischhoff, 1977; Maulida et al., 2015) State that the category of item difficulty level depends on the skills of the research subject and the general skills of all research subjects (not only individuals).

Relevant

The Testlet model assessment instrument integrated with the E-ujian website to measure students' HOTS was relevant, with an average percentage of 93.65 % for all indicators measured. The indicators measured in the relevant characteristics are (1) the substance of the questions refers to the scope of the

competency achievement indicators, and (2) the questions instruments can interpret the cognitive domain. (Safitri, 2020) explains that preparing an assessment instrument that follows the competence will allow the alignment of the instrument form, material, and time allocation required with the depth of the material.

Representative

The questionnaire analysis results showed that the assessment instrument developed represented all the material taught. The percentage obtained was within the excellent criteria (96% for all indicators at all stages). Recapitulation of the questionnaire responses results can be seen in Table 4.12. The results of this analysis are also supported by the analysis of the material experts at the validation stage. The quality of the material in the material expert validation sheet contained the accuracy and truth indicators.

Impractical

The practicality based on the validation of the assessment experts showed that the developed assessment instrument was declared highly feasible with a validation percentage of 95 %. The assessment instrument was easy to grade, making it easier for teachers to correct so that an objective and fast assessment is produced and minimizes student cheating during the test. Students also get a discussion at the end of the assessment session. The weakness of the developed assessment instrument was that it was a pay-to-use platform, so it is not appropriate if used for formative assessments, such as daily tests. Besides, the camera feature caused slow website access in large quantities and numbers. At the same time, the website often experienced a decline in function. Also, students at junior high schools were not familiar with the website, which affected the implementation of the test instrument.

The Testlet model assessment instrument integrated with the E-ujian website is considered practical if it is easy to use. This aligns with (Arifin, 2012) statement, which explains that practical means are easy to use. The practicality of an assessment instrument is seen not only from the manufacturer's side but also by people who want to use the instrument. Based on the questionnaire results, the developed product was declared impractical because it has many weaknesses compared to its advantages.

Specific

The questionnaire analysis of student and teacher responses shows that the Testlet model assessment instrument was specific within excellent criteria. The average percentage was 91%. The assessment instrument was specific because it had questions that focused on the science object material and observations. Another specific characteristic of the Testlet model assessment instrument is its usability and characteristics that distinguish it from other instruments. The Testlet model assessment instrument developed can measure junior high school students' HOTS.

Students' HOTS Profile

The students' HOTS profile data was obtained from the implementation results in the large-scale trial. The measured HOTS aspects consisted of five skills, namely: (1) critical thinking skills, (2) creative thinking skills, (3) problem-solving skills, (4) decision-making skills, and (5) argumentative skills. A recapitulation of the number of items for each HOTS aspect indicator can be seen in Table 1. The Testlet model questions developed in this study were adapted from (Shiell & Slepkov, 2015), an integrated set of questions consisting of three or more items in the form of multiple tiered choices. The items in the Testlet model test consisted of main questions and supporting questions. The questions were made to provide information to other

supporting questions. One of the questions on the Testlet model assessment instrument that can measure students' HOTS is in the main item number 1, which can be seen in Table 3.

Table 3. Sample Question

Main Questions
Yahya is experimenting with the following steps: 1. Cutting a piece of tissue paper with 4 x 12 cm. 2. Draw a line with a red marker 2 cm from tissue paper. 3. Take a measuring cup and fill it with water as high as 1 cm. 4. Dip the paper into a measuring cup filled with water. 5. Observe and record the changes.
Integration 1 1.1 The steps taken in scientific investigations on the main problem are… A. Observation because it uses the five senses and produces data. B. Measurement because it uses a measuring instrument in a ruler to measure length. C. Making Inference because it is intended to find patterns or relationships between aspects to be observed. D. Communicating because it tells about the data to be obtained. 1.2 The correct hypothesis for the results of experiments carried out on the main problem is... A. The water will rise and wet the tissue paper. B. The water will rise, causing the marker lines to fade. C. The water will rise, causing the line stain to turn orange. D. The water will rise, causing the line stain to turn pink. 1.3 The best way to convey experimental results on the main problem is…… A. Make a comparison graph of the length of the markers before and after. B. Displaying experimental tissue paper directly. C. Create a table containing colors formed from markers.

Every main question contains three supporting questions that are related to each other. In the supporting questions, some indicators can measure the HOTS aspects of students. Item 1.1 measures the HOTS aspect in the form of the skills to argue with the indicator to be measured, namely Data. Item 1.2 measures the HOTS aspect of critical thinking skills with indicators in the form of combining information. Item 1.3 measures the aspect of problem-solving skills with indicators of compiling problem-solving. Therefore, there are two to three aspects of HOTS. The data from the large-scale trial analysis on 32 seventh-grade junior high school students related to the students' HOTS profiles are presented in Figure 2.

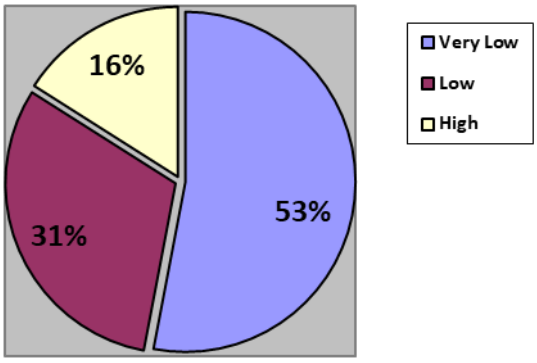

Figure 2. The Percentage of Student HOTS Aspect Profiles

The students' HOTS profiles were divided into three categories. There were 53 % of students in the very low HOTS category, 31 % of students in the low

HOTS category, and 16% in the high HOTS category. Based on the results of data analysis on proportional characteristics, the questions developed were in the difficult category.

Achievement of HOTS Aspects for Students with High HOTS

There were 16 % students out of 32 students belonging to the high HOTS category. The students in this category were familiar with the level of difficulty of the questions on the Testlet model assessment instrument. This fact is supported by the results of student questionnaire analysis in the very good HOTS category of 75 %. Questionnaire items for student responses contain statements stating that students are familiar with the difficulty of the questions being worked on. The high HOTS of students is also due to the understanding of concepts, application and use, utilization, and selection of operations that have reached the excellent category. This fact is coherent with the results of the questionnaire analysis, which shows the percentage of achievement in that aspect is 75 % and 70 %. The highest aspect is the skills to argue with a percentage of 3.6 %. On the other hand, the lowest aspect is the skills to make decisions at 2.7 %.

The percentage of critical thinking skills was still in the very low category because the average achievement for each question with the HOTS indicator was only 24 %. Furthermore, the aspect of decision-making skills was 17 % in the very low category. In the argumentative skills, the indicators consisted of (1) claim, (2) data, and (3) rebuttal. 80 % of students in the high HOTS category correctly answered questions by following the data indicator. In the rebuttal indicator, only 40 % of students answered correctly. Lastly, in the claim indicator, 60 % of students answered correctly.

Achievement of HOTS Aspects for Students with Low HOTS

The numbers of students in the low HOTS category were ten students. The percentage of achievement in this category was 31 %. The highest percentage was 48 % in this category, and the lowest percentage was 26 %. The students' low HOTS was caused by unfamiliarity with working on HOTS questions, as shown by the results of the questionnaire analysis on statement 6, which obtained a percentage of 29 % in the poor category. This result is coherent with (Anggraini et al., 2019; Rahayuningsih & Jayanti, 2019). They state that students could not complete HOTS because of difficulty determining ideas to explain them. This research is also supported by (Heong et al., 2012; Yee et al., 2015) that difficulties in generating ideas experienced by students will cause them to experience technical problems in completing their assignments. Another factor is the unfamiliarity with the HOTS questions. (Hall & Piazza, 2008; Lavy, 2020; Rochman & Hartoyo, 2018) state that other factors influencing students' low HOTS are culture and character. Different cultures and characters will affect the students' mindset.

For students with low HOTS, the order of achievement of the highest HOTS aspects to the lowest is (1) critical thinking skills (7.6 %), (2) problem-solving skills (6.2 %), (3) decision-making skills (6.2 %), (4) argumentative skills (5.8 %), and (5) creative thinking skills (5.2 %). Critical thinking skills were the aspect that got the highest achievement, although it was still classified in the very low criteria (25 %). (Febriana & Sinaga, 2021; Samsul Hadi et al., 2018; Priyadi et al., 2018) claim in their research that the students' low critical thinking skills are caused by several things, including (1) students had difficulty in completing and answering the questions given and (2) students had

difficulty in connecting the results of calculations with the phenomena presented. This claim is supported by the results of the questionnaire analysis on statement items number 9 and 10 which obtained a percentage of 33 % and 37 % in the poor category.

The Achievement of HOTS Aspects of Students with poor HOTS

The percentage of students in the poor HOTS category was 53 %. This aspect was the most dominant than other HOTS categories. The aspect that obtained the highest percentage of achievement was the critical thinking skills, with the highest percentage of 13.2 % and the lowest percentage of 7 %. The lowest aspect was the creative thinking skills. At a glance, the results for students with poor HOTS were similar to those of students with low HOTS. However, the difference lies in the percentage of decision-making and argumentative skills. For students with poor HOTS, argumentative skills were higher than decision-making skills. On the other hand, students in the low HOTS category had higher decision-making skills than argumentative.

Another difference was the achievement of the creative thinking aspect. The percentage of achievement for students in the very low category was 13 %, while for students in the low HOTS category was 17 %. The achievement percentage in both categories was poor because it was only 25 %. For students in the poor HOTS category, their creative thinking skills were lower than those in the low HOTS category.

Low creative thinking skills can be caused because students are unfamiliar with questions that train creative thinking skills. Besides, the learning methods do not train students' creative thinking skills (Rochman & Hartoyo, 2018). Creative thinking skills can be trained by improving learning strategies and models. This argument is supported by (Mahanal et al., 2019; Rohim & Susanto, 2012; Saputri et al., 2019), who state that the improvement of learning strategies is to change the learning model that can facilitate communication between students and students and teachers and students. Also, to foster students' creative thinking skills.

CONCLUSION

The Testlet model assessment instrument integrated with the E-ujian website is feasible for measuring the junior high school students' HOTS. The developed assessment instrument measures students' HOTS based on several characteristics, namely (1) valid, (2) reliable, (3) discriminatory, (4) disproportionate, (5) relevant, (6) representative, 7) impractical because it has many obstacles when implementing it, and (8) specifically functions to measure students' HOTS with aspects measured in the form of critical thinking skills, creative thinking skills, problem-solving skills, decision-making skills, and argumentative skills; specifically for the object of science.

This research can be used as a reference for teachers or students of the educational program as one of the innovations of learning assessment instruments that can be applied during the Covid-19 pandemic. The development of the Testlet model assessment instrument can also be used to measure the five skills of HOTS: (1) critical thinking skills, (2) creative thinking skills, (3) problem-solving skills, (4) argumentative skills, and (5) decision-making skills.

REFERENCES

Anggraini, N. P., Budiyono, & Pratiwi, H. (2019). Analysis of higher order thinking skills students at junior high school in Surakarta. *Journal of Physics: Conference Series*, *1211*(1). https://doi.org/10.1088/1742-6596/1211/1/012077

Arif, S., & Yuhdi, A. (2020). Integration

of high order thinking skills in research method subject in university. *Britain International of Linguistics Arts and Education (BIoLAE) Journal*, *2*(1), 378–383. https://doi.org/10.33258/biolae.v2i1.207

Arifin, Z. (2012). *Evaluasi pembelajaran* (Edisi Revi). Direktorat Jendral Pendidikan Islam Kementrian Agama RI.

Arikunto, S. (2013). *Dasar-dasar evaluasi pendidikan*. Bumi Aksara.

Arnellis, A., Fauzan, A., Arnawa, I. M., & Yerizon, Y. (2020). The effect of realistic mathematics education approach oriented Higher order thinking skills to achievements' calculus. *Journal of Physics: Conference Series*, *1554*(1). https://doi.org/10.1088/1742-6596/1554/1/012033

Asrizal, A., Amran, A., Ananda, A., & Festiyed, F. (2018). Effectiveness of adaptive contextual learning model of integrated science by integrating digital age literacy on grade VIII students. *IOP Conference Series: Materials Science and Engineering*, *335*, 012067. https://doi.org/10.1088/1757-899X/335/1/012067

Azmi, N. L., Nurhayati, S., Priatmoko, S., & Wardani, S. (2021). Pengembangan instrumen tes untuk mengukur HOTS peserta didik pada materi laju reaksi. *Chemistry in Education*, *10*(1), 45–52.

Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research*, *2*(2), 189–193.

Damayanti, I. (2017). Pengembangan instrumen penilaian testlet untuk mengukur kemampuan berpikir tingkat tinggi pada materi elektrokimia untuk siswa SMK. *Jurnal Pendidikan IPA*, *8*(1), 57–79.

Dharmawati, Rahayu, S., & Manahal, S. (2016). Pengembangan instrumen asesmen berpikir kritis untuk siswa smp kelas vii paa materi interaksi makhluk hidup dengan lingkungan. *Jurnal Pendidikan*, *1*(8), 1598–1606. https://doi.org/10.17977/jp.v1i8.6677

Ennis, R. (2011). Critical thinking: Reflection and perspective Part II. *Inquiry: Critical Thinking Across the Disciplines*, *26*(1), 19–28. https://doi.org/10.5840/inquiryctnews20112614

Fahmina, S. S., Masykuri, M., Ramadhani, D. G., & Yamtinah, S. (2019). Content validity uses Rasch model on computerized testlet instrument to measure chemical literacy capabilities. *AIP Conference Proceedings*. https://doi.org/10.1063/1.5139755

Febriana, R., & Sinaga, P. (2021). Evaluation of critical thinking skills of class x high school students on the material of Newton's laws. *Journal of Physics: Conference Series*, *1806*(1). https://doi.org/10.1088/1742-6596/1806/1/012012

Hadi, Samsul, Retnawati, H., Munadi, S., Apino, E., & Wulandari, N. F. (2018). The difficulties of high school students in solving higher-order thinking skills problems. *Problems of Education in the 21st Century*, *76*(4), 97–106.

Hadi, Syamsul, & Novaliyosi. (2019). TIMSS Indonesia (Trends in International Mathematics and Science Study). *Prosiding Seminar Nasional & Call For Papers Program Studi Magister Pendidikan Matematika Universitas Siliwangi*, 562–569.

Halim, A., Ngadimin, Soewarno, Sabaruddin, & Susanna. (2018). Improvement of high order thinking skill of physics student to prepare human resources in order to faced of

global competition in asean economic community. *Journal of Physics: Conference Series*, *1116*(3). https://doi.org/10.1088/1742-6596/1116/3/032009

Hall, L. A., & Piazza, S. V. (2008). Critically reading texts: What students do and how teachers can help. *The Reading Teacher*, *62*(1), 32–41. https://doi.org/10.1598/rt.62.1.4

Hayashi, P., Abib, G., & Hoppen, N. (2019). Validity in qualitative research: A processual approach. *Qualitative Report*, *24*(1), 98–112. https://doi.org/10.46743/2160-3715/2019.3443

Heong, Y. M., Yunos, J. M., Othman, W., Hassan, R., Kiong, T. T., & Mohamad, M. M. (2012). The needs analysis of learning higher order thinking skills for generating ideas. *Procedia - Social and Behavioral Sciences*, *59*, 197–203. https://doi.org/10.1016/j.sbspro.2012.09.265

Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2018). Developing of Computerized adaptive testing to measure physics higher order thinking skills of senior high school students and its feasibility of use. *European Journal of Educational Research*, *9*(1), 91–101. https://doi.org/10.12973/eu-jer.7.3.555

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of item response theory item fit indices for the graded response model. *Organizational Research Methods*, *14*(1), 10–23. https://doi.org/10.1177/1094428109350930

Lavy, S. (2020). A review of character strengths interventions in twenty-first-century schools: Their importance and how they can be fostered. *Applied Research in Quality of Life*, *15*(2), 573–596. https://doi.org/10.1007/s11482-018-9700-6

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183. https://doi.org/10.1016/0030-5073(77)90001-0

Lutviana, E., Rahardjo, S. B., Susanti, E., Yamtinah, S., Mulyani, S., & Saputro, S. (2019). The computer-assisted testlet assessment instrument to measure students' learning difficulties in chemical bonding. *Journal of Physics: Conference Series*, *1156*(1), 1–5. https://doi.org/10.1088/1742-6596/1156/1/012019

Mahanal, S., Zubaidah, S., Sumiati, I. D., Sari, T. M., & Ismirawati, N. (2019). Ricosre: A learning model to develop critical thinking skills for students with different academic abilities. *International Journal of Instruction*, *12*(2), 417–434.

Masitoh, L. F., & Aedi, W. G. (2020). Pengembangan instrumen asesmen Higher order thinking skills (HOTS) matematika Di SMP kelas VII. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, *4*(2), 886–897.

Mateucci, M., & Stracqualursi, L. (2006). Student assessment via graded response model. *Statistica*, *66*(4), 435–447.

Maulida, Muhibuddin, & Yuzrizal. (2015). Analisis indeks kesukaran dalam pengembangan item tes pada konsep sel tingkat sekolah menengah atas. *Jurnal Edubio Tropika*, *3*(1), 42–45.

Momani, R. T. (2017). Using item response theory to evaluate self-directed learning readiness scale. *Journal of Educational and Developmental Psychology*, *8*(1), 14. https://doi.org/10.5539/jedp.v8n1p14

Morse, B. J., Johanson, G. A., & Griffeth, R. W. (2012). Using the graded

response model to control spurious interactions in moderated multiple regression. *Applied Psychological Measurement*, *36*(2), 122–146. https://doi.org/10.1177/0146621612438725

Muna, A., Noer, A. M., & Linda, R. (2017). *The development of instrument diagnostic test as a testlet model for learning difficulties detection on chemistry (buffers)*. Universitas Riau.

Nisa, A. K. N., Widyastuti, R., & Hamid, A. (2018). Pengembangan instrumen assesment higher order thinking skill (HOTS) pada lembar kerja peserta didik kelas VII SMP. *Prosiding Seminar Nasional Matematika Dan Pendidikan Matematika UIN Raden Intan Lampung*, *1*(2), 543–556.

Nova, R. A., Parno, & H Koes Supriyono. (2016). Pengembangan instrumen asesmen penguasaan konsep tes testlet pada materi suhu dan kalor. *Jurnal Pendidikan: Teori, Penelitian Dan Pengembangan*, *1*, 1197–1203.

Nurtanto, M., Kholifah, N., Masek, A., Sudira, P., & Samsudin, A. (2021). Crucial problems in arranged the lesson plan of vocational teacher. *International Journal of Evaluation and Research in Education (IJERE)*, *10*(1), 345. https://doi.org/10.11591/ijere.v10i1.20604

Polya, G. (1993). *How to solve it*. Princenton University Press.

Prayitno, A., & Nofiana, M. (2020). Pengembangan instrumen evaluasi high order thinking skills pada materi jaringan hewan dengan bentuk two-tier multiple choice question. *Jurnal Pendidikan Biologi*, *1*(1), 1–11.

Priyadi, R., Mustajab, A., Zaky Tatsar, M., & Kusairi, S. (2018). Analisis kemampuan berpikir kritis siswa SMA kelas X MIPA dalam pembelajaran fisika. *Jurnal Pendidikan Fisika Tadaluko Online (JPFT)*, *6*(1), 53–55.

Qi, S., & Mitchell, R. E. (2012). Large-scale academic achievement testing of deaf and hard-of-hearing students: Past, present, and future. *Journal of Deaf Studies and Deaf Education*, *17*(1), 1–18. https://doi.org/10.1093/deafed/enr028

Rahayuningsih, S., & Jayanti, R. (2019). High order thinking skills (HOTS) mahasiswa program studi pendidikan matematika dalam menyelesaikan masalah grup. *Majamath: Jurnal Matematika Dan Pendidikan Matematika*, *2*(2), 88–93.

Retnawati, H., Djidu, H., Kartianom, Apino, E., & Anazifa, R. D. (2018). Teachers' knowledge about higher-order thinking skills and its learning strategy. *Problems of Education in the 21st Century*, *76*(2), 215–230. https://doi.org/10.33225/pec/18.76.215

Rochman, S., & Hartoyo, Z. (2018). Analisis high order thinking skills (HOTS) taksonomi menganalisis permasalahan fisika. *Science and Physics Education Journal (SPEJ)*, *1*(2), 78–88. https://doi.org/10.31539/spej.v1i2.268

Rohim, F., & Susanto, H. (2012). Penerapan model discovery terbimbing pada pembelajaran fisika untuk meningkatkan kemampuan berpikir kreatif. *UPEJ Unnes Physics Education Journal*, *1*(1).

Rusilowati, A., Kurniawati, L., Nugroho, S. E., & Widiyatmoko, A. (2016). Developing an instrument of scientific literacy assessment on the cycle theme. *International Journal of Environmental and Science Education*, *11*(12), 5718–5727.

Safitri, R. K. (2020). *Pengembangan tes diagnostik model testlet untuk mendeteksi kesulitan belajar peserta didik pada materi suhu dan perubahannya*. Universitas Negeri

Semarang.

Saputri, A. C., Sajidan, Rinanto, Y., Afandi, & Prasetyanti, N. M. (2019). Improving students' critical thinking skills in cell-metabolism learning using stimulating higher order thinking skills model. *International Journal of Instruction*, *12*(1), 327–342. https://doi.org/10.29333/iji.2019.12122a

Shiell, R. C., & Slepkov, A. D. (2015). Integrated testlets: A new form of expert-student collaborative testing. *Collected Essays on Learning and Teaching*, *8*(2), 201–210.

Silver, E. . (1997). Fostering creativity through instruction rich in mathematical problem solving and problem posing. *Zentralblatt Für Didaktik Der Mathematik*, *29*, 75–80.

Slepkov, A. D. (2013). Integrated testlets and the immediate feedback assessment technique. *American Journal of Physics*, *81*(10), 782–791.

Slepkov, A. D., & Shiell, R. C. (2014). Comparison of integrated testlet and constructed-response question formats. *Physical Review Special Topics - Physics Education Research*, *10*(2), 1–15. https://doi.org/10.1103/PhysRevSTPER.10.020120

Suek, L. A. (2021). Item analysis of an english summative test. *PEJLaC: Pattimura Excellence Journal of Language and Culture*, *1*(1), 9–18. https://doi.org/10.30598/pejlac.v1.i1.pp9-18

Tan, S. Y., & Halili, S. H. (2015). Effective teaching of higher order thinking (HOT) in education. *The Online Journal of Distance Education and E-Learning*, *3*(2), 41–47.

Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. ERIC. https://doi.org/10.1016/0022-4405(76)90066-2

Toulmin, & Stephen, R. (1979). *An Introduction to Reasoning*. Machmillan.

Utomo, D. W., & Kustijono, R. (2015). Pengembangan Sistem Ujian online soal pilihan ganda dengan menggunakan software wondershare quiz creator. *Jurnal Inovasi Pendidikan Fisika (JIPF)*, *4*(3), 1–6.

Wahyuni, I. T., Yamtinah, S., & Budi, T. (2015). Pengembangan instrumen pendeteksi kesulitan belajar kimia kelas x menggunakan model testlet. *Jurnal Pendidikan Kimia*, *4*(1), 222–231.

Wang, Y., & Ruhe, G. (2007). The cognitive process of decision making. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *1*(2), 73–85. https://doi.org/10.4018/jcini.2007040105

Widana, I. W. (2018). Higher order thinking skills assessment towards critical thinking on mathematics lesson. *International Journal of Social Sciences and Humanities (IJSSH)*, *2*(1), 24–32. https://doi.org/10.29332/ijssh.v2n1.74

Yee, M. H., Yunos, J. M., Othman, W., Hassan, R., Tee, T. K., & Mohamad, M. M. (2015). Disparity of learning styles and higher order thinking skills among technical students. *Procedia - Social and Behavioral Sciences*, *204*, 143–152. https://doi.org/10.1016/j.sbspro.2015.08.127