



Development of CBT- Based Assessment Instruments Using WQC Application to Measure HOTS in Impulse Momentum

Anggun Wulandari*, Agus Suyatna, Viyanti Viyanti, Undang Rosidin

Departement of Physics Education, Lampung University, Indonesia

*e-mail: anggunwulandari1998@gmail.com

Received: Maret 17, 2021

Accepted: June 31, 2021

Published: July 19, 2021

Abstract: This study aims to develop an instrument for HOTS questions with the help of the Wondershare Quiz Creator (WQC) application on Momentum and Impulse materials with design and material feasibility, suitable settings for HOTS questions with four types of HOTS questions, namely Multiple Choice, Multiple Responses, Sequence and Matching. The research method used is Research and Development (R & D). This type of research is used to adapt the research design by Borg & Gall. The product development stages consist of research and data collection, planning, product development, product validation, product revision, and product results. The data collection technique used expert validation in developing HOTS questions based on CBT, then the data were analyzed quantitatively and descriptively. The analysis shows that the validity of the design has an average value of 4.21 with very high quality or validity and the validity of the material has an average value of 3.75 with high quality or valid. The highest scores for design and material validity were multiple responses 4.22 and 3.79. At the levels on C4, C5, and C6, there is no difference in the validity and complexity of thinking caused by different types of questions.

Keywords: CBT, HOTS, Momentum and Impulse, Wondershare quiz creator

DOI: <http://dx.doi.org/10.23960/jpf.v9.n1.202110>

INTRODUCTION

The quality of education is a problem that the government has always strived to improve. One way to control quality in education is to conduct an assessment (Sutama, Sandi, and Fuandi, 2017: 106). Assessment is a process of collecting and processing information to measure the achievement of student learning outcomes. Mardapi (2008) suggests that assessment is an aspect that determines the quality of education. To improve the quality of education, efforts are needed both in terms of process and results. One indicator that can be improved is student learning outcomes from time to time. Learning outcomes can be obtained from the evaluation of learning conducted by teachers on students.

To find out the learning outcomes of students, the teacher must conduct an assessment that produces information about the achievement of competencies that have been possessed by students. In carrying out the assessment, teachers and education units must refer to the Regulation of the Minister of Education and Culture number 23 of 2016 concerning Educational Assessment Standards, namely, attitude assessment, knowledge assessment and skills assessment (Betty, 2017). In fact, the implementation of Indonesia's national education has not been fulfilled properly, this is shown from the achievement of education in Indonesia which is still not encouraging. The results of *The Program for International Student Assessment (PISA) study* on scientific literacy skills ranked 64th out of 65 countries in 2012, and in 2015 Indonesia ranked 64th out of 72 participating countries, in 2018 Indonesia ranked 6th. 43 of the 79 participating countries. Likewise, Indonesia's creativity index only 0.20, or ranks 115th out of 139 countries (Martin Prosperity Institute, 2015). The same thing also happened to Indonesia's *Global index of cognitive skills and educational attainment* which had a position of $z = -1.84$ – the lowest of the 40 countries tested (*The Learning Curve*, 2014).

According to (Budiman.A, Gilani, 2014) u ntuk support the communication skills, critical thinking and creative learners with the teachers can do about that characterized melatihkan HOTS. In this case, of course, the teacher must look for more weighty material references. The problem faced by teachers is that the ability of teachers to develop HOTS assessment instruments is still lacking, besides that there is no assessment instrument specifically designed to train HOTS so it is necessary to develop a HOTS assessment instrument.

Along with technological advances in all fields including education, the demand for mastery of ICT is a must, including in the evaluation of learning. Along with the development of technology, learning evaluation has shifted from paper-based (manual) to computer-based, of course, to reduce the weaknesses of manual learning evaluation and realize *paperless* in the digital era. Teachers as educators are required to provide effective and interesting learning evaluation instruments so that students are interested in continuing to learn and practice, and can be used in teaching and learning activities both inside and outside the classroom for all levels of education. Learning evaluation instruments can be in the form of visual, audio, audio visual, multimedia and others. One example of a multimedia-based and appropriate learning evaluation instrument in the current era is the *Computer Based Test (CBT)*.

Software that can be used in developing computer-based test assessment instruments, namely, *Edmodo*, *Quipper School*, *Wondershare Quiz Creator*, and so on. In this case, the researcher will develop HOTS questions using *Wondershare Quiz Creator (WQC)*. WQC is one of the *software* or *software* that can be used to make questions, quizzes or tests online (Haida, 2017). The use of this *software* is very *user friendly* or *familiar* and very easy to use so there is no need for a complicated programming language (Jayanta, 2013). This *software* can be used to create various types of questions and different cognitive levels.

Several researchers have tried to examine the use of WQC in previous studies. Khaldun (2019) examines the development of HOTS chemistry questions using WQC for the form of *Multiple Choice* questions. In his research, the quality of the questions is very feasible and good for use in the evaluation process, thus the use of this application is very effective. Selvi (2017) developed an online daily test assessment instrument to measure material mastery on physics material. In terms of product manufacturing, the product composes multiple choice questions in the cognitive domain C1-C3 using the *Hypertext Preprocessor software application* and the product is made suitable for use.

The research conducted by Arinil (2019) is the development of a daily test assessment instrument using the *wondersaher quiz creator* on statistics material for class XII SMA. The form of questions developed in this study are *multiple choice* and *short essay*. The daily test instrument using WQC on statistical material is valid, reliable and effective. Based on the research above, the three researchers have developed products by utilizing various software applications. However, the assessment instrument product developed only focuses on multiple choice questions. From various software applications used by previous researchers, researchers are interested in using WQC software applications. *The software* has the convenience of making questions or quizzes because it does not require expertise in programming languages.

For this reason, it is necessary to have an innovative CBT model HOTS assessment with appropriate types of questions to stimulate students' higher-order thinking skills in Physics subjects by applying basic competencies and indicators and having the characteristics of good test instruments used for assessment according to the dimensions of knowledge with needs analysis in the field. The difference with previous studies is related to the question instruments made and the form of the questions to be developed. The question instrument that will be made is the HOTS question and the form of the questions contained in the previous research only focuses on the form of *multiple choice* (*multiple choice*) and *short essay* (*short essay*). However, the form of questions that will be developed in this study are *multiple choice*, *multiple response*, *matching* and *sequence*. Researchers want to know how the types of questions are valid for measuring HOTS using WQC. In addition, the researcher also wanted to find out whether the CBT setting in accordance with using WQC to measure HOTS in a test question could affect students' higher order thinking skills. And whether between types of questions there are differences in the level of validity, complexity and level of thinking.

METHOD

This research is research and development (*Research and Development R&D*). The product developed in this study uses the Borg & Gall (2003) model which consists of 10 development steps. In this study, the researcher used 6 steps consisting of: (1) research

and information gathering, (2) planning, (3) initial product development, (4) product validation, (5) initial product revision, (6) final product.

The procedure for developing an analysis of the form of questions to measure HOTS on Momentum and Impulse material can be seen in the figure:

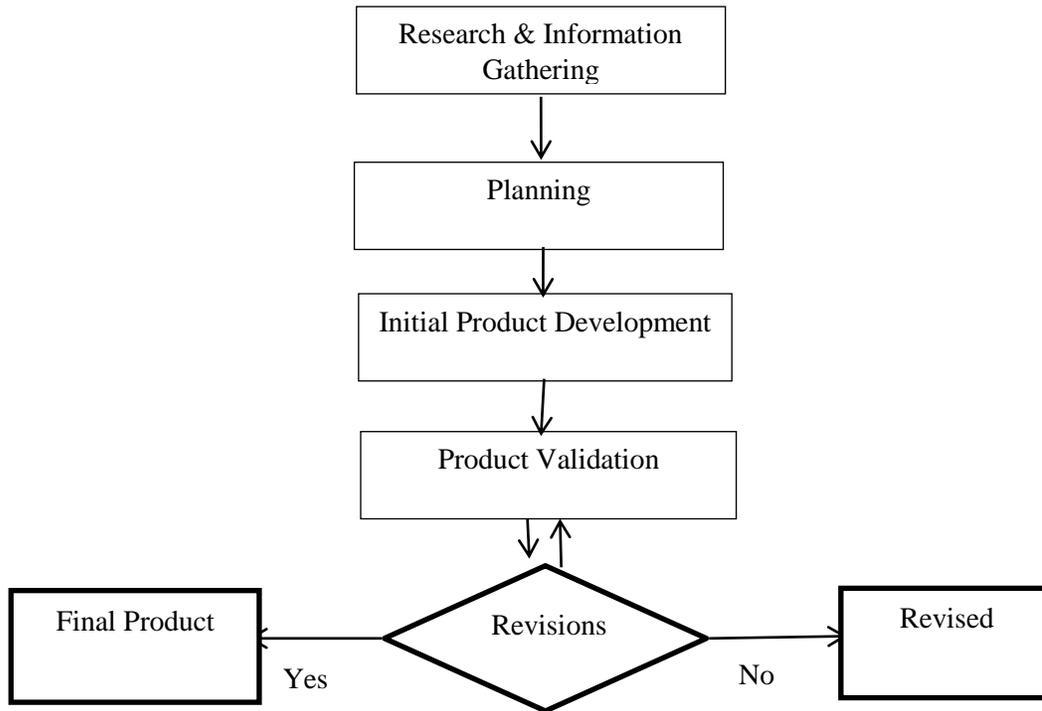


Figure 1. Product Development Procedure

Data was collected by means of literature study, namely reading literature from books, journals, and articles. The information obtained includes the formulation of learning objectives that can be used as the basis for the formulation of test questions. In this study, the researcher first chose the basic competencies that could be made *HOTS* questions. Researchers must first choose KD which can be made *HOTS* questions. Should choose KD that contains KKO which refers to the cognitive domains C4, C5 and C6. Based on the results of the literature study, this will be a reference for researchers to develop *prototype* products that are in accordance with the form of questions to measure *HOTS* on Momentum and Impulse material. The planning stage is by compiling a *HOTS* question grid in the form of product design using the help of the *WQC* application. The question instrument consists of 4 types of questions, namely *multiple choice*, *multiple response*, *sequence*, and *matching* on momentum and impulse material. The product development stage is carried out by formulating a stimulus. Stimulus used can be in the form of pictures, graphs, tables, discourses, and videos. Next, write *HOTS* questions using the *WQC application*. Setting the *WQC* which consists of setting time, random, answer subtion, feedback and setting scoring and weighting scores. The resulting development is a test instrument with a variety of questions that can be used to measure *HOTS*. Validation Phase After developing the product, the next step is the validity test conducted by a team of experts. This validity test was carried out by three lecturers and two teachers who are experts in physics with master qualifications in physics education. This repair was carried out according to the advice of a team of experts. The final product is the result of research and development. The results are in the form of a prototype of the *HOTS* questions for Momentum and Impulse. The validity test has been carried out by a team of experts, then the test results are valid for use.

Research Design & Procedures

Data collection techniques used expert validation sheets and respondent questionnaires. Expert validation involved several experts to evaluate the initial product developed by the researcher. The data used is in the form of a validation sheet given to the expert. The validation sheet is used to collect data in the form of responses and suggestions as a basis for revising the initial product. The data obtained from the experts are discussed as a reference for revising the product until it is declared feasible to be tested. This validation includes material, construction, and language validation

Data Collection and Instrument

Expert Validity

The activities carried out in this stage are analyzing the results of the validator's assessment of the material validation sheet, construction and CBT settings that have been made by the researcher. To make it easier to analyze the data from the validation results, the activities carried out are: (Rachma, 2018). Recapitulate all validator statements into a table which includes: Aspects of assessment (A_i), criteria (K_i) and validator research results (V_{ij}). The average of each criterion from all validators can be found by using

$$K_i = \frac{\sum_{j=i}^n V_{ji}}{n}$$

The information K_i is average of the i criteria, V_{ji} is the score of the results of the j th validator research for the i -th criterion, n = Number of validators.

The average total validator of all criteria is searched by the formula:

$$RTV_i = \frac{\sum_{j=i}^n A_i}{n}$$

The information RTV_i is average total validation, A_i is average aspect of i , and n is number of validators. The results obtained are then written in the appropriate table column.

Then compare the total average with the validity indicators according to Widoyoko (2009), which is in Table 1

Table 1 . Media/Material Eligibility Criteria

Score Interval	Validity Interval
$RTV \geq 4.20$	Very good
$3.40 \leq RTV < 4.20$	Good
$2.60 \leq RTV < 3.40$	Pretty good
$1.80 \leq RTV < 2.60$	Not good
$RTV < 1.80$	Not good

(Widoyoko, 2009)

The results of the analysis of the data obtained are used as a reference for revising the assessment media. The media is said to be valid if the score obtained from the validator is at least good enough. If the validation results do not meet these categories, then the media needs to be revised to meet the minimum categories. Differences in validity and complexity of thinking caused by different types of questions are obtained by conducting a one-way ANOVA test with the following conditions. If the value of Sig > 0.05, there is no difference in average validity and complexity caused by different types of questions. If the value of Sig < 0.05, there is a difference in average validity and complexity caused by different types of questions.

RESULT AND DISCUSSION

The results of the development (*Development*)

CBT-based HOTS development products can be seen in Figure 2 below.

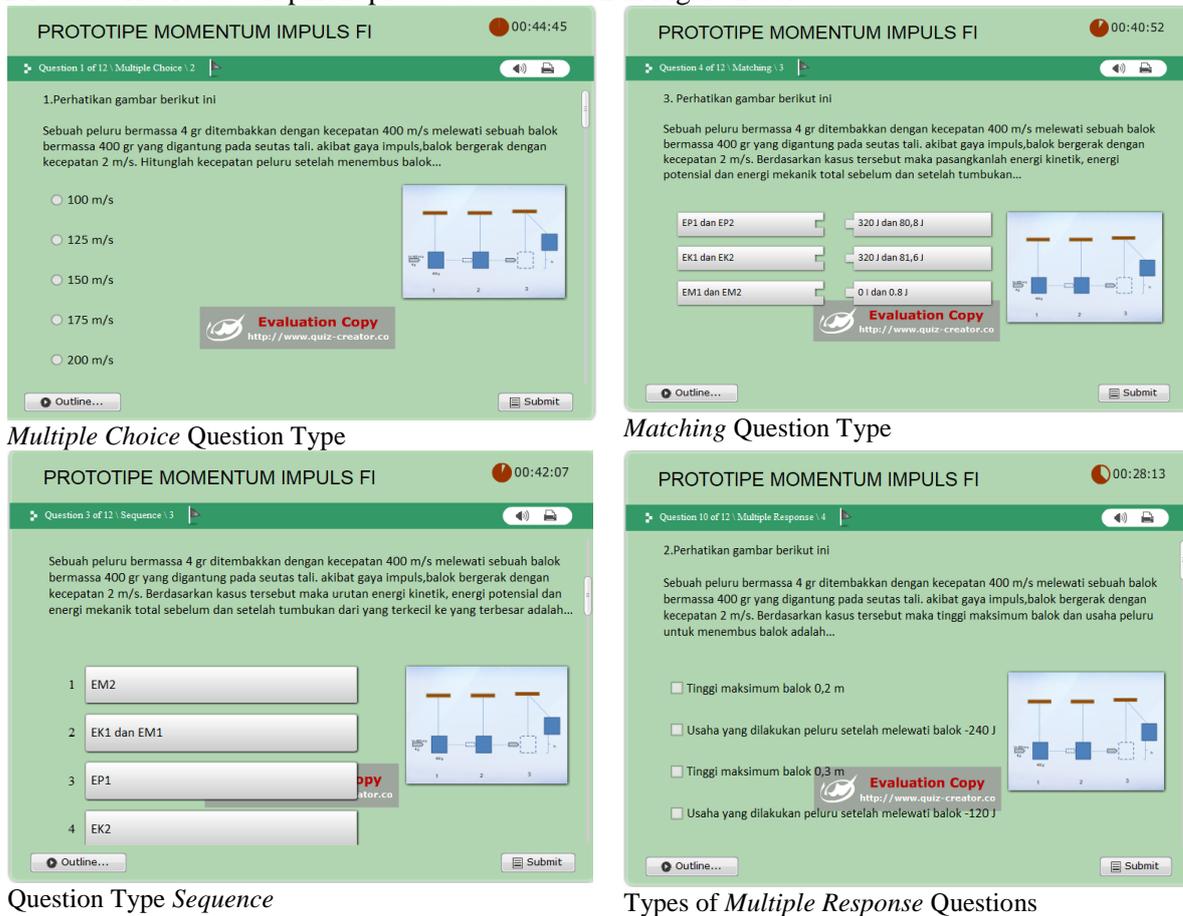


Figure 2. CBT-Based HOTS Development Products

Validation Test

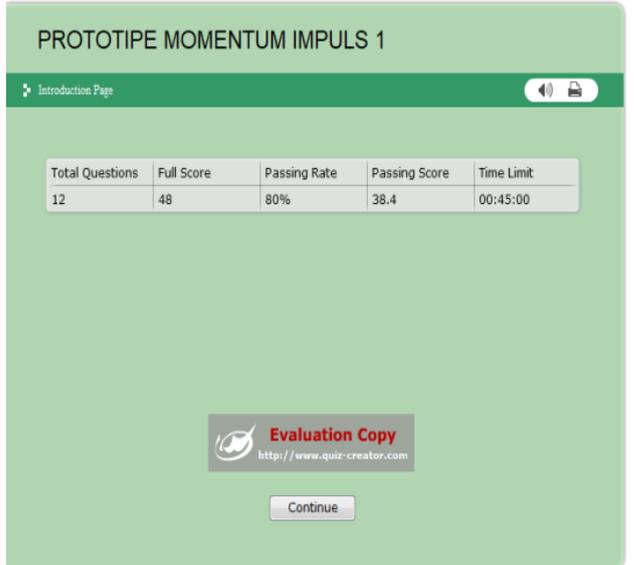
To validate the development of the CBT-based HOTS assessment on momentum and impulse materials, two aspects were validated, namely design and material validation. The overall results of the average validation of the validators can be seen in the table below.

Table 2. Instrument Validation Results.

Criteria	Aspect					Average	Note:
	Theory	Construction	Language	Question Type	CBT settings		
C4	3.75	3.88	4.50	3.95	4.12	4.04	Good
C5	3.78	4.08	4.50	3.95	4.08	4.07	Good
C6	3.70	4.28	4.50	3.95	4.12	4.11	Good
Overall Average Tital						4.07	Good

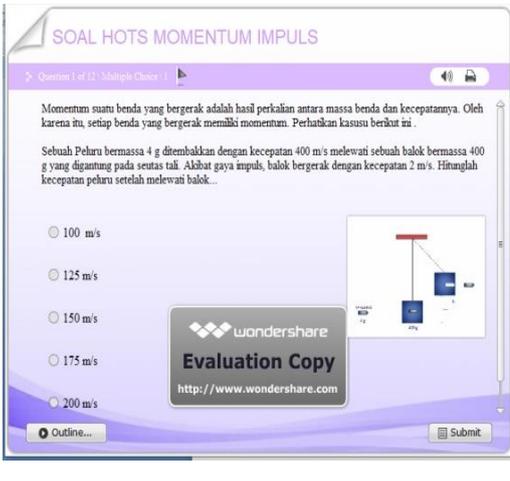
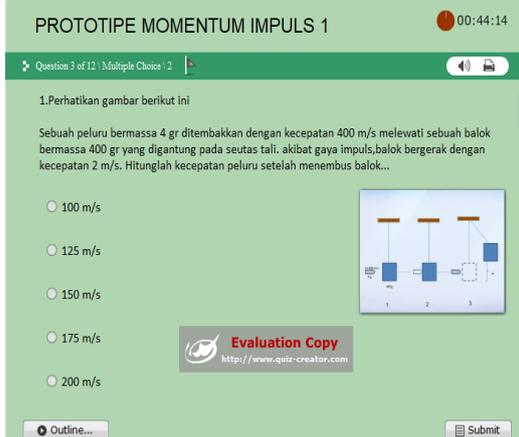
From the validation results above, it shows that the average value of the validity is in the "good" category. So that the results of the assessment are obtained from the formula for finding the average of each aspect assessed, in the table above there are five aspects that are assessed, namely material, construction, language, type of questions, and CBT settings. The overall results of the CBT-based HOTS instrument validation assessment on the momentum and impulse material from the validator is 4.07, meaning that the materials, constructions, language and illustrations that have been designed by the researcher are valid. Although the expert has stated that the product developed is valid and can be used for research, previously the expert also provided criticism and suggestions. Suggestions for improvement from each validator for the design test are presented in Table 3

Table 3. Design test suggestions and improvements

Suggestions and Improvements	Corrective action
<p>The question doesn't have time yet</p> 	<p>Fixing questions by adding time according to what students need to do it</p> 

Suggestions for improvement from each validator for material testing are presented in Table 4.

Table 4. Design test suggestions and improvements

<p>The level of homogeneity of the questions is less logical, causing misconceptions</p> 	<p>Replacing Stimulus questions that are less logical so that it is clear what you want about the question and does not cause misconceptions mis</p> 
--	---

The test results of the momentum and impulse prototype design experts are presented in Table 5

Table 5. Design Expert Test Accumulation Analysis Results

Question Type	Average Score	Quality Statement
MC	4.20	Valid /Very Good
MR	4.22	Valid /Very Good
S	4.20	Valid /Very Good
M	4.22	Valid /Very Good

Based on the results of the design expert test, it can be seen in the table that for the four types of questions a score of > 4.20 with a statement of quality is Valid / Very Good. This means that these four types of questions are declared valid for use. Expert test results The momentum and impulse prototype materials are presented in Table 6.

Table 6. Material Expert Test Accumulation Analysis Results

Question Type	Average Score	Quality Statement
MC	3.75	Valid /Good
MR	3.79	Valid /Good
S	3.75	V a lid /Good
M	3.73	Valid /Good

Based on the table of the results of the validity of the material test, it can be seen that the prototype made is valid for use. It can be seen from the average score for each type of question. The results of the accumulation of material expert tests for the type of

multiple choice questions with an average score of 3.75, multiple response 3.79, sequence 3.75 and matching 3.73.

Table 7. Data on Differences in Validity and Complexity Levels Between Question Types and Thinking Levels

No	Question Type	Thinking Level	Average Validity	Average Complexity
1	Multiple Choice	C4	4.01	3.80
2		C5	4.06	3.80
3		C6	3.92	3.80
4	MultipleResponse	C4	4.06	4.00
5		C5	4.07	4.00
6		C6	3.93	4.00
7	Sequence	C4	4.05	4.00
8		C5	4.08	4.00
9		C6	3.87	4.00
10	Matching	C4	3.99	4.00
11		C5	4.09	4.00
12		C6	3.93	4.00

The data above is used to find differences in the level of complexity and validity between types of questions and levels of thinking, so the researchers used SPSS and tested *one way ANOVA* with a *test of homogeneity of variances*. The results of the *one way ANOVA test* using SPSS to measure the level of complexity and validity can be seen in Table 8

Table 8. Test results of Homogeneity of Variances
Test of Homogeneity of Variances

Score	Levene Statistics	df1	df2	Sig.
	.440	1	22	.514

Table 9 . One Way Anova Results

ANOVA					
Score	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.018	1	.018	2,625	.119
Within Groups	.152	22	.007		
Total	.170	23			

Based on the output results of the *Test of Homogeneity of Variances* in Table 8, a significance value of 0.514 is obtained. Therefore, the significance value is greater than 0.05, so H_0 is accepted, which means that the four samples (test results) have homogeneous variances. Based on the results of the *Anova* output in Table 9, a

significance value of 0.119 is obtained. Because the significance value is greater than 0.05, then H_0 is accepted, which means rejecting H_1 . This result shows that there is no difference in the average validity of thinking and thinking complexity caused by the types of *multiple choice, multiple response, sequence and matching questions*.

The results of setting the HOTS questions for Momentum and impulse in the WQC application are presented in Table 10

Table 10. Setting HOTS Questions on the WQC Application

Question Settings	Information
Question Settings	
<i>Question Properties</i>	<ul style="list-style-type: none"> • If the student answers correctly then the score is given according to the type and level of the question. • If the answer is wrong = 0
<i>Feedback</i>	Feedback will be given when students finish answering one question. <ul style="list-style-type: none"> • If the answer to the question is wrong, the feedback will provide a key so that students can recall. • If the answer is correct, a statement will be given to support the answer.
Quiz Settings	
<i>Time Limit</i>	The time limit given is according to the level of difficulty of the questions
<i>Randomization</i>	Randomization was carried out for all questions
<i>Answer Submission</i>	<ul style="list-style-type: none"> • incorrect/correct answers are given to each item directly when students answer • tell the correct answer when students answer • give cross for wrong answer • students can repeat answers when working on all questions

The results of *setting the HOTS* questions on impulse momentum with the *WQC* application in the table explain that this HOTS question product is valid using any type of form in the application settings. This is in line with (Pranata, 2020) Time management is considered good in limiting the work on questions. According to the expert, time settings can be provided for each question according to the level of difficulty and the process of working on the problem. Random setting of questions and answers is considered good in preventing cheating by students when taking tests. The answer submission setting is considered good for stimulating HOTS, the feedback setting is considered good by the expert. Likewise in setting scoring da x n weighting of questions that have been adapted to the cognitive and the level of difficulty about it rated well by all the experts and practitioners in stimulating HOTS.

Types of questions that are valid in measuring HOTS on momentum and impulse materials using WQC

Based on the results on the validity of the question design, it shows the feasibility of the questions to be used by students. The results of the analysis of the HOTS test of momentum and impulse material using WQC can be seen in Table 5 . From the data presented, the results of the analysis of the design expert test show that the four types of questions with quality are valid to use. Material validity tests the suitability of the content of the material and the language used in the questions. The results of the analysis of the material expert test on the HOTS material on momentum and impulse using WQC can be seen in Table 6 . Based on the data presented in Table 6 , the four types of questions with good categories and quality are valid for use. Based on the results of the analysis of the design expert test and the material expert test conducted, it was stated that the four types used were valid to measure HOTS in line with the opinion (Rahmani, 2015) that the questions were said to be valid or have high validity, which are questions that can measure the expected competencies. While questions that are invalid or have low validity mean that the questions cannot measure the expected competencies.

Based on the data, it is known that in the results of the analysis of the design expert test and the material expert test, the type of question that gets the highest average score is the *Multiple Response* question type and the second is the *matching* question type , this is also in line with Eka's statement (2020) the possible types of HOTS questions that suitable to be developed for CBT is by filling in the blanks, matching, and sequences and reinforced by Pranata's statement (2020) the type of matching question is suitable as a type of question that is suitable for measuring students' HOTS because in the type of matching question students must match all questions and answer choices with right. However, if students match questions with answers, there is one answer that is not correct with the question, then students will not get a score.

Barratt, (2014) *Higher Order Thinking Skill (HOTS)* is a skill that demands creative, critical, analytical thinking on data and information in solving a problem. The form of this type of question requires the test taker to choose two possible answers. The forms of possible answers that are often used here are true and false or yes and no. According to Rahmawati (2017) *Multiple Response*, to make multiple-choice questions with multiple answers (more than one correct answer). There are several advantages of this type of question form: (a) it can measure various levels of cognitive ability, (b) it can cover a wide scope of material, and (c) it can be scored easily, quickly, and objectively.

The average score was obtained from five validators consisting of three physics education lecturers at the University of Lampung and two physics subject teachers with a master's qualification in physics education. Based on these results, it means that the four types of valid questions are used on the product to measure the HOTS of momentum and impulse material using the WQC application. An instrument is said to be valid if it has analysis results in accordance with predetermined criteria. This agrees with (Pranata , 2020) that the types of questions true or false, multiple choice, fill in the blank and matching have a suitable level of eligibility to stimulate HOTS.

Based on the results of the validity, the developed HOTS instrument meets the valid category, because the aspects of the developed instrument have an average value of 4 which is in the valid category. This value is obtained from the results of the assessment carried out by the validator on the product that has been developed in the form of the HOTS instrument by making several revisions to obtain an instrument that is ready to be tested. Based on this, it is supported by the data from the validity results that have been validated by 5 validators, it can be seen that it is included in the "Good" category, thus the instrument that has been made has been declared valid.

Appropriate Settings for Measuring Material Momentum and Impulse HOTS in WQC Applications

There are two settings for the *questions*, namely *question settings* and *quiz settings*. *Question settings* consist of *question properties* and feedback. Question properties, namely the value of each question if they answer true / false, if they answer correctly students will get different scores between questions, this is seen from the type and level of the question.

Feedback will appear when students have answered questions correctly or incorrectly, feedback that appears when true/false is different, namely when true, a statement will appear that strengthens the answer, whereas if students answer incorrectly then feedback will appear a statement that makes students recall the lesson that. This makes it easier for students to be able to see their mistakes in line with (Rahmawati, 2017) For students, tests using computers are more interesting and teachers do not need a lot of time to correct student test results. The CBT/computer-based test directly provides feedback, which means that the computer itself will correct students' assignments. Quiz settings consist of time limit, randomization, and answer submission.

The time limit is the time limit given to each question, the time limit for each question is different, according to the length of time students answer the questions given, and this is seen from the level of difficulty of the questions. The total time given is 45 minutes and in accordance with the practicality test, the time used is in accordance with the time provided. Randomization is randomization of questions when the questions are done, this randomization is found in the answers to the questions. *Answer submission* is a wrong/correct answer, in this question the answer is given directly to the item, not at the end. The settings used are declared valid so that with these settings they can measure HOTS questions on momentum and impulse using the WQC application on four different types. This is in line with Eka's statement (2020) Based on the results of the CBT-HOTS analysis, it is necessary to design student CBT-HOTS HOTS according to the needs in the field including the design of quiz settings, design of setting questions, design of time, and design of stimulus contained in several types questions, namely filling in the blanks, matching, and sequences. Reinforced by the statement (Pranata, 2020) that timing, *random* questions and answers, *feedback*, *answer submission*, scoring and weighting are assessed both by experts and practitioners.

The Effect of Different Types of Questions and Thinking Levels Based on the Average Results (mean)

The one way ANOVA test to measure the difference in validity and complexity between types of questions and levels of thinking presented in Table 7 explains that at the thinking levels C4, C5, C6 there are no differences in validity scores and thinking complexity caused by different types of questions, as well as in the one-way ANOVA test. the sig result is $0.119 > 0.05$. This is because the average validity is not too far away or not too influential, thus using any type of question will remain valid. In line with Harvianita's statement (2020) at the level of thinking C4, C5, and C6, there is no difference in validity scores or in thinking complexity caused by different types of questions, meaning that using any type of question remains valid.

At the level of thinking C4 type *multiple response questions* obtained the highest average validity so that it is more valid to use compared to other types. Thinking level C5 there is no difference in validity scores and thinking complexity, but for the type of *matching* question, the average validity is the highest. In line with Pranata (2020) the type of matching question is suitable as a type of question that is suitable for measuring students' HOTS because in the type of matching question students must match all questions and answer choices correctly. However, if students match questions with answers, there is one answer that is not correct with the question, then students will not get a score. But at the level of thinking C6 for the average validity with *multiple response* and *matching* question types, it gets a higher value so that it is more valid than other types, on the type of question. Seen from Table 7, the relationship between the level of thinking and the complexity of thinking is correlated, which means that the higher the level of thinking, the higher the complexity of thinking.

CONCLUSION

The valid Momentum and Impulse question instruments to measure HOTS in this study, namely all types of questions used including *multiple responses*, *multiple choice*, *sequences*, and *matching* with the help of the *Wondershare Quiz Creator* (WQC) application were declared valid. Setting CBT The *WQC*-based test instrument to measure *HOTS* on the momentum and impulse materials developed is declared valid, this is in accordance with the results of the questionnaire test. The appropriate setting for the test instrument is to adjust the student's score according to the weight and provide feedback on each question. The time limit given is 45 minutes by activating random questions. Based on the results of the study, it can be seen that at the level of thinking C4, C5, C6, there is no difference in validity scores or thinking complexity due to different types of questions.

REFERENCES

Arinil, Mega, 2019. *Development of Daily Test Assessment Instruments Using Wondershare Quiz Creator on Statistics XII High School Material*. Downloaded from <http://digilib.uinsby.ac.id/id/eprint/30267> on 20 January 2020

- Barratt, Caroline. 2014. *Higher Order Thinking And Assessment*. International Seminar on current issues in Primary Education : PGSD Study Program, University of Muhammadiyah Makassar
- Betty, Asrotaja. 2017. *Development of a Higher Order Thinking (HOT) Evaluation Instrument Based on a Computer Based Test (CBT) in the Sensory System Sub Material for Class XI High School Students* (thesis). University of Surabaya. Surabaya
- Budiman, A., Jailani. 2014. Development of Higher Order Thinking Skills (HOTS) Assessment Instruments in Mathematics Subjects for SMP Class VIII Semester 1. *Journal of Mathematics Education Research* , 1(2).
- Eka, Diah P., Suyatna, Agus., Viyanti, V. 2020. Design of Computer-Based Tests for Higher Order Thinking Skills in Static Fluid Materials. *J. Phys. Conf.Ser* . **1467** 012059.
- Haida, Dafitri. 2017. "Utilization of Wondershare Quiz Creator in Computer-Based Tests", *Journal of Information Systems* , 1(1):8–18 .
- Jayanta, Profit. 2013. "Development of Online Mathematics Tests with Dynamically Adjusted Difficulty Levels for Fifth Grade Students of SD Laboratorium Singaraja", *e-Journal of the Graduate Program of Ganesha University of Education* , vol 3, 2013, 7.
- Khaldun., L. Hanum, SD Utami. 2019. Development of Computer-Based Higher Order Thinking Skills Chemistry Questions with Wondershare Quiz Creator Materials for Hydrolysis of Salts and Buffer Solutions. *Indonesian Journal of Science Education (Indonesian Journal of Science Education)*. 7(2), pp. 132-142, 2019
- Mardapi. 2008. *Techniques for Preparation of Test and Non-Test Instruments* . Scholar Partners, Yogyakarta. 67.
- Pranata, Bayu., Suyatna, Agus., Rosidin, Invite. 2020. Development of Computer Based Test (CBT) High Order Thinking Skill (HOTS) on Electromagnetic Induction Material. *JP3I (Indonesian Journal of Psychology and Education)* , 9(2), 2020, 83-98.
- Rachma, Hanun Nur. 2018. *Development of Gamelan (Adventure Math Game) as a Media for Student Daily Tests Based on Story Problems*. (Thesis). UIN Sunan Ampel Surabaya.Surabaya.
- Rahmani, MN, Kurnia and Nurdini A. 2015. Analysis of the Quality of Question Items Made by a Biology Teacher Class X SMA Negeri 1 Tanah Pinoh. *Journal of Education and Learning*. 4(2):1-16.
- Rahmawati, Diah As'ari. 2017. Utilization of Wordshare Quiz Creator in Making Arabic Questions. *Journal of Arabic Studies* , 2 (1), 2017: 37-46. State University of Malang
- Selvi, Anggraini. 2017. *Development of Online Daily Test Assessment Instruments To Measure Mastery Of Physics Materials And To Know The Learning Responses Of High School Students* . (Thesis). Yogyakarta State University. Yogyakarta .
- Sutama, Sajidan & Afandi. 2017. *Development of Science Learning Models to Empower Higher-Level Thinking Skills*. Proceedings of the National Seminar on Science Education (SNPS) Sebelas Maret University
- The Learning Curve. 2014. Index- *Which countries have the best schools ?*
<http://thelearningcurve.pearson.com/index/index-ranking>

Widoyoko, Eko Putro. 2009. *Evaluation of the Learning Program*. Yogyakarta : Pustaka Pelajar.