IJOLAE Indonesian Journal on Learning and Advanced Education

http://journals.ums.ac.id/index.php/ijolae

Education Application Testing Perspective to Empower Students' Higher Order Thinking Skills Related to The Concept of Adaptive Learning Media

Hernawan Sulistyanto[1], Sofyan Anif[2], Sutama[3], Sabar Narimo[4], Anam Sutopo[5], Muhammad Izzul Haq[6], Gamal Abdul Nasir Zakaria[7]

[1-5]Faculty of Teacher Training and Education, Universitas Muhammadiyah Surakarta, Indonesia
[6]Faculty of Arts, McGill University, Canada
[7]Sultan Hasanal Bolkiah Institute of Education, Universiti Brunei Darussalam, Brunei Darussalam

DOI: 10.23917/ijolae.v4i3.19432
Received: August 3rd, 2022. Revised: August 25th, 2022. Accepted: September 2nd, 2022
Available Online: September 5th, 2022. Published Regularly: September 1st, 2022

Abstract

This article aims at arguing for the importance of the testing step when designing an educational application by taking a case study from the development of adaptive learning media. The media contains a set of instruments that are specifically built to empower students' critical thinking skills. Three aspects that are considered in testing this educational application are application validity at each stage of system development, measurement of the final system feasibility test for user needs, and system implementation by running learning media on the test sample. Implementation of testing on application products is carried out according to system requirements and models. The existence of the characteristics of adaptive media and the diversity of menus in the application implies the importance of doing a lot of improvisation when carrying out tests, such as determining the right test cases, choosing the appropriate test model and method, determining a suitable test environment, and considering several other aspects aimed at optimizing test results. obtained in order to ensure the quality of learning media products. This study analyzed the test data using Likert scale as an interpretation of the results of the validation assessment from the experts by referring to certain perceived standards of assessment. Meanwhile, the analysis of the data from the feasibility test results from a sample of 20 students using the system usability scale (SUS) instrument. The technique to test the effectiveness was using a pretest-posttest control group design with a sample of 98 students. Parametric/non-parametric data analysis was then applied to analyze the data on the results of testing the effectiveness or efficacy of adaptive media products in improving students' higher order thinking skills (HOTS). Based on the testing steps applied to the application of adaptive learning media, the results obtained that the product was considered feasible and effective in empowering students' HOTS. The study concludes that the educational application testing that has been carried out is able to provide an objective and independent view of the application of adaptive learning media which will be useful in operational functions to understand the level of effectiveness in its implementation before being widely used in learning.

Keywords: adaptive learning media, education application, HOTS, measurement, system development

Corresponding Author:
Hernawan Sulistyanto, Faculty of Teacher Training and Education, Universitas Muhammadiyah Surakarta
Email: hernawan.sulistyanto@ums.ac.id

1. Introduction

Educational application testing is an in-depth investigation carried out to obtain information about the quality of a learning media product being tested (Maulana, A., et. All., 2020). The increased visibility of educational applications as system elements and the "costs" arising from application failures have motivated good planning through careful and accurate testing. This makes testing educational applications an important stage in the development of learning media. The reasons why testing is necessary are application developers are not good enough programmers; application developers may not be able to concentrate specifically on avoiding mistakes; application developers sometimes forget to use structured programming in full; application developers are sometimes bad at doing things; and application developers must can distinguish what other developers or users are saying and what they really think (Schwan et al., 2018).

Currently, learning aids are being developed in the form of information technology-based learning media (Afandi et al., 2018) (Seechaliao, 2017). Various advanced concepts and algorithms have been implanted as the embodiment of learning applications that follow today's technological developments. Therefore, it is important to adhere to the correct testing rules in maintaining the quality of the resulting media products.

Testing can be done by evaluating the application configuration consisting of requirements specifications, design descriptions and the resulting program (Kurniawan, D., et. all., 2022) (Purmadi & Surjono, 2016). The evaluation results are then compared with the expected test results. If errors are found, the application must be repaired and then tested again. So, basically testing activities can be considered destructive rather than constructive. However, the importance of testing the application of learning media and its implications for quality cannot be overemphasized because it involves a series of production activities where the chances of human error can be very large. Therefore, the development of educational applications should be accompanied by quality assurance activities (Bedjou et al., 2015).

One of the products produced in this research is adaptive learning media. The adaptive concept is pinned to describe the media's intelligence in adjusting the presentation of material according to the character of student learning. In this article, it will be explained how the form of testing that has been carried out during research in the implementation of learning media validation, measuring application feasibility, and evaluating product effectiveness in HOTS empowerment.

2. Method

This study includes two main aspects, namely empirical studies and practice testing steps in the development of adaptive learning media. The study presented in this article is the result of a study of several literature sources, both printed and electronic. Sources include primary and secondary sources that were studied empirically and descriptively. Furthermore, the testing steps were thoroughly practiced in the development of adaptive learning media based on the identification of student learning characters. The learning media development method used Research and Development (R&D) (Kusuma et al., 2017) with the appropriate Luther development model (Sulistyanto, H. et al., 2019). By following the steps in the R&D method, testing was carried out.

Testing activities consisted of two stages of the R&D method, namely the development stage consisting of validation and feasibility testing of product drafts and the testing stage called the application effectiveness test step. In the feasibility test, the product draft was evaluated three times to nine students to study the products and assess the performance of the products for improvement s. In the feasibility test, suggestions for improvement were obtained for the product draft for revision before beng developed further.

Feasibility tests were also conducted by referring to the standard software application design testing using the System Usability Scale (SUS) model (Sauro, 2011). The feasibility test was conducted by 20 students as suggested by Roscoe (Sugiyono, 2017). The validation and feasibility test steps at the development stage are conducted with the following procedures (Jingyun & Takahiko, 2015): (1) Determining the test targets and product draft test subjects, namely learning technology experts, media experts, learning style experts, and student users. Experts are asked to provide input related to the design of adaptive learning media, namely several main aspects of learning media and the truth and accuracy of content management that is developed in accordance with existing indicators; (2) The validation test was carried out by two graphic design experts for application media, linguists, learning style experts, and learning technology experts. The product was handed over to experts to provide reviews in the form of suggestions and input based on their expertise (Wang & Mendori, 2016). The results were analyzed and then used as the basis for conducting initial revisions before proceeding to the feasibility test and implementation in the field; (3) Conduct a feasibility test by involving 20 students in the first feasibility test. This test aimed at determining the feasibility of using the product.

Based on the suggestions, inputs, and improvements, the results of the analysis of this activity became the basis for conducting the final revision. The second and third feasibility tests were conducted in small groups of five students as a descriptive and meaningful test of the feasibility of media products before field testing; (4) The evaluation process was also carried out by teachers and students as respondents after using media products. The effectiveness evaluation was conducted in the form of pre-test and post-test to measure students' understanding ability or cognitive learning outcomes. The effectiveness analysis technique employed a pretest-posttest control group design research design as described in Table 1 below.

Table 1. Design of Application Media Product Testing

Group	Pre-Test	Treatment	Post-Test
Re	T_1	X	T_2
Rc	T_3	-	T_4

R : experimental and control groups taken by random clusters
T_1, T_3 : pre-test
X : treatment with product
T_2, T_4 : post-test (Leow & Neo, 2014)

In testing the effectiveness of the product used a population of all students. A sample of 98 students was taken by using cluster random sampling method to determine which students were members of the experimental class and the control class. The form of the developed application product is shown by the system diagram in Figure 1 below.

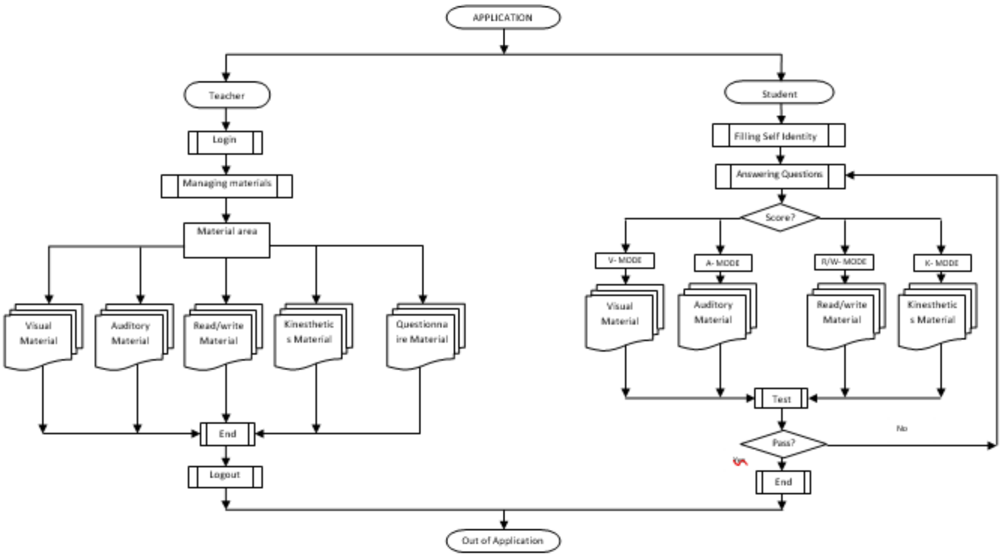

Figure 1. Developed Application Product

3. Results and Discussion

The benchmark used to judge that the test is good according to Pressman (2010) and Sommerville (2011) is that the test has a high probability of finding errors. For this to happen, testers must understand how the app can fail. Each test conducted must have a different purpose. Furthermore, the test conducted is the best type of test. Testing is possible in several ways. The test used was the one having the greatest probability of uncovering all categories of errors (with the least amount of time and effort). So, the test was complex.

Product draft validation was a process to assess whether the developed product draft in accordance with the existing theoretical requirements. This validation was rational, as it was based on facts in the field. The interpretation of the validation assessment category for learning media products in this study fell into the following scale (Wu & Leung, 2017):

0%-20%	: Strongly Unvalidated
21%-40%	: Unvalidated
41%-60%	: Fairly Validated
61%-80%	: Validated
81%-100%	: Strongly Validated

The product drafts were in the form of adaptive application media, model guides, lecturer guides, and student guides validated by experts. Experts who validated the media applications were those with the division of the field of validation as shown in Table 2 below.

Table 2. List of Validation Fields on Adaptive Media Applications

Num	Validation fields	Number of validators
1	Adaptive media learning concept	2 experts
2	Learning technologies	2 experts
3	App design and graphics	2 experts
4	Determination of learning style preferences	2 experts
5	Model guides and learning modules	2 experts
6	Implementation of the learning concept	2 experts

Table 3. The Results of The Validation of The Adaptive Media Learning Model

Num	Aspects	Interpretation Index (%)
1	Syntax	87,50
2	Social system	87,50
3	Reaction principle	75,00
4	Support system	75,00
5	Learning impact	87,50
	Interpretation index average	**82,50**

The summary of expert validation on the concept of adaptive learning media syntax is shown in Table 3. The results of expert validation obtained an average value of 82.50% interpretation which means the concept of the adaptive learning media model is very feasible. The syntax of the adaptive learning media is considered to have clear, systematic, logical stages, and can be used to measure critical and creative thinking skills. Students were judged to be able to formulate answers in their own words, motivated to ask questions and active in debate when applying the model. The model is considered capable of providing opportunities for students to take the initiative, be responsive, innovative, communicative and respectful of each other.

The model is also equipped with learning tools in the form of modules that are considered good by experts. Students as targets of model application, can understand the material, are able to work together, and empower critical thinking skills as the instructional impact of model application. Experts assess this adaptive learning model can provide an accompaniment impact, namely familiarizing students in solving problems actively and establishing good communication between students so that in the end it can empower critical thinking skills. The advice given by the expert is that the support system is better with a little more detail.

The summary of expert validation or learning technology experts on adaptive learning applications can be seen in Table 4 below.

Table 4. Summary of Validation Results by Learning Technology experts

Num	Aspects	Number of indicators	Interpretation Index (%)
1	Teaching materials	4 indicators	81,25
2	Learning process	3 indicators	91,67
3	Assessments	4 indicators	81,25
4	Learning support activities	3 indicators	91,67
5	Materials for increasing student competence	5 indicators	85,00
	Interpretation index average		**86,00**

Table 4 presents information on the results of the assessment of two learning technology experts with an average index of interpretation of 86.00%. Based on the rating scale category, it can be seen that the two learning technology experts considered it very feasible to draft an adaptive learning application product that was developed to be piloted. Suggestions and inputs given by learning technology experts 1 and 2 can be concluded as follows: 1) In general, the material presented needs to be slightly improved on the depth and breadth of the material. 2) In the aspect of assessment, a more even distribution of material is needed; 3) need to clarify the description of indicators and try to add or adjust the time allocation. 4) The suitability of the question with the make elaboration problem aspect is slightly clarified.

Furthermore, a summary of the validation of the application design and graphic experts is shown in Table 5 below. Table 5 provides information on the results of the assessment of two design and graphic software application experts with an average interpretation index of 83.13%.

Table 5. Summary of validation results by design and application graphic design experts

Num	Aspects	Number of indicators	Grade Point Average
1	Organization	4 indicators	87,50
2	Attractiveness	2 indicators	87,5
3	Font shape and size	3 indicators	91,67
4	Space (blank space)	1 indicator	100
5	Consistency	3 indicators	79,17
6	Image/video/text presentation	5 indicators	92,50
7	Language	2 indicators	81,25
	Interpretation index average		**83,13**

Based on the rating scale category, it can be seen that both design and graphic experts rated it very feasible for the draft adaptive learning application to be developed. Suggestions for improvement from design and graphic experts include improvements to: 1) arrangement of manuscripts, pictures, illustrations to make them easier to understand; 2) content needs to pay attention to free space; 3) consistency of letter shape and font size on each page from beginning to end: 4) linguistic aspects of the sentences used are made more flexible so that they are easy to understand; 5) management of free space in each view is optimized with appropriate content.

Data analysis of respondents' responses in the first feasibility test used a standard instrument of web application feasibility test, namely the System Usability Scale (SUS) (Martins et al., 2015). The summary of the data analysis of the feasibility test results is shown in Table 6 below.

Table 6. Summary of the results of the first feasibility response with SUS

Number of respondents	Score result count (sh)	Value (sh x 2,5)
40	1242	3105
	Average	**77,625**

Through the calculation process according to the SUS rules in (Sauro, 2011) above, the final result of the average feasibility score is 77.625 as shown in Table 6.

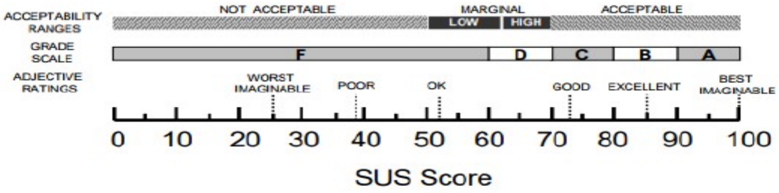

Figure 2. Categories of SUS assessment results (Sauro, 2011)

Based on Figure 2, it can finally be determined that the average respondent's feasibility test results are worth 77.625 which is greater than the SUS average value of 68. These results indicate that there is no problem with the application made. Based on Figure 2, the value of 77.625 is in the good rating domain, class C scale, and at an acceptable interval. This shows that the application of adaptive learning media is feasible and can be used in learning.

The next stage is testing the effectiveness of adaptive learning media. The first step is to test the balance between the experimental class and the control class. The data used are the results of the pre-test scores of the two groups. Because the results of the prerequisite analysis found that the data were not normally distributed, the non-parametric Wilcoxon Signed Rank Test (paired test) and Mann Whitney U Test (unpaired test) were used (Vong & Kaewurai, 2017). The results of the analysis are shown in Table 7 below.

Table 7 The Results of The Analysis of the Balance Test Between Experimental and Control Groups

Test Statistics	
	Pretest
Mann-Whitney U	1118.500
Wilcoxon W	2393.500
Z	-.920
Asymp. Sig. (2-tailed)	.358

a. Grouping Variable: Group

Based on the results of the non-parametric analysis as presented in Table 7, the comparison of student scores in the experimental group and the control group obtained a pre-test score with had a significance value of 0.358. As the value of Sig = 0.358 > 0.05, it can be concluded that there is no significant difference between the experimental group and the control group. This means that in the two groups there is no difference in initial ability before the treatment in the experimental group. The test results using the Wilcoxon Signed Rank Test is shown in Table 8 below.

Table 8. Results of Non-Parametric Analysis of Wilcoxon Signed Rank Test on The Experimental and Control Groups

Test Statistics		
	Pretest Experiment - Posttest Experiment	Pretest Control-Posttest Control
Z	-6.186^b	-6.197^b
Asymp. Sig. (2-tailed)	.000	.000

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

Based on the results of the non-parametric analysis of the Wilcoxon Signed Rank Test as presented in Table 8 above, the comparison of the pre-test and post-test scores in the experimental group was 0.000. Because the value of Sig = 0.000 < 0.05, the difference between the pre-test and post-test mean values in the experimental group was significant. Meanwhile, the comparison of pre-test and post-test scores in the control group was also 0.000. Because the value of Sig = 0.000 < 0.05, the difference in the average value between the pretest and post-test in the control group was significant. It suggested that both in the experimental group and the control group, the pretest and post-test scores were different. This means that based on the pre-test to post-test scores, students from both groups had the same ability to change

Furthermore, the results of the test on the post-test scores between the two groups are shown in Table 4.35 below. Based on Table 4.35, the significance = 0.000 <0.05 so it means that the results of the post-test scores between the two groups have differences due to the treatment in the experimental group, namely learning using applications.

Table 9. Results of the Analysis of Post-Test Scores between Experimental and Control Groups

Test Statistics	
	Post-test
Mann-Whitney U	34.000
Wilcoxon W	1309.000
Z	-8.425
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: Group

The analysis was also conducted on the increase in the difference between the results of the pre-test and post-test (gain) as an indicator of the effectiveness of adaptive media used in learning. The description of the gain score is shown in Table 10 below.

Table 10. The Results of the Descriptive Analysis of the Gain Scores of the Two Groups' Test Results

Num	Group	*Gain score (%)*		
		Min	Max	Average
1	Control	40,91	73,68	55,11
2	Experiment	68,00	95,83	86,05

In accordance with the interpretation category (Hake, 1999) the average gain score obtained in the control class is 55.11%, so belongs to less effective. While in the experimental class, the score of 86.05% belongs to the effective category.

The description of the gain score per aspect of the HOTS is shown in Table 11 below.

Table 11. Description of the Gain Score on the Aspect of Critical Thinking Skills

Group	HOTS aspects					
	Interpretation	Inference	Explanation	Analysis	Evaluation	Self-regulation
Experiment	93.18	76,74	96,33	97,87	63,01	98,13
Control	65,90	44,52	71,68	64,83	36,91	66,34

The gain score obtained in the HOTS aspect in the control class has a range between 36.91% to 71.68% with an average of 58.37% in the sufficient category.

Meanwhile, for the experimental class, the gain scores ranged from 63.01% to 98.13% with an average of 87.54% in the effective category.

The conclusion obtained is that in the HOTS aspect the increase in gain score in the experimental class is greater than the control class which indicates that adaptive learning media is effective in empowering students' HOTS.

Educational application testing is the process of running and evaluating learning media software manually or automatically to assess whether the application meets the requirements or not (Khodadi & Abadeh, 2016) (Clune and Rood, 2011) (Yoshii & Nakajima, 2012) (Nakagawa and Maldonado, 2011). In short, testing is an activity to find and determine the difference between the expected results and the actual results.

Application tests follows the building of applications from abstract concepts of user needs. The purpose of testing is to "disassemble" applications built. According to (Chang et al., 2015), Jin and Xue (2011) and Kumamoto et al. (2010) testing intends to find errors in the application compiler program and evaluate its quality. The purpose of application testing according to (Xie et al., 2015) is to assess whether the application developed has met user needs, assess whether the application development stages are in accordance with the methodology used, and make documentation of test results that inform the suitability of the application product assessed to the existing. The data collected during the test gave a good indication of the overall reliability and quality of the application.

According to Pressman (2010), application program testing has several important objectives, namely (1) testing is carried out with the intention of finding errors; (2) test success is the ability to find errors that have never been found before; and (3) a good test case is a test case that has a high probability of finding errors that have never been found before. Objectivity in testing can be achieved if there are several actors involved during the test, including according to Lamas et al. (2013) namely the customer (the team that contracts the application developer), the user (the group that will use the application), the application developer (the team that builds the application), and the application testing team (a special team assigned to test the functionality in the software application). In addition, it should always be based on the principle that testing can be traced to customer needs, testing should be planned before testing, testing should start with small results and then move on to larger things, over-testing will not be possible, and testing should be carried out by a third party (Baoling et al., 2020) (Jiang and Lu, 2012) (Lemos et al., 2011).

The implementation of learning media application testing usually matches to the development methodology used. Reza (2010) and Sommerville (2011) stated that testing is done after the programming stage but testing planning has been conducted from the analysis stage. Overall, the stages in testing include determining what will be measured, how the test will be conducted, building a test case, which is a set of data or situations that will be used in testing, then determining the expected results or actual results, running the case. test and compare the test results with the expected results.

The analysis phase emphasizes the validation of user requirements to ensure that the requirements have been correctly defined.

The purpose of testing at this stage is to obtain a feasible requirement and to ascertain whether the need is formulated properly. The testing factors conducted at the analysis stage are requirements related to the methodology, determining functional specifications, determining usability specifications, determining portability requirements, and determining system interfaces. Design phase testing aims to evaluate the structure of the software that comes from the needs. General needs are dissected into more specific forms. The testing factors conducted at the design stage are design related to requirements, suitability of design with methodology and theory, design portability, design maintenance, correctness of design related to function and data flow, and completeness of interface design. Testing at the implementation stage is a test of the units made before being integrated into the overall application. The testing factors conducted at this stage are data integrity control, program correctness, ease of use, and development of operating procedures.

Referring to Wen-hong and Xin (2010), engineered products can be assessed by: (1) knowing the specific function that the product is designed to perform. Tests are conducted to ensure that each function is fully operational and to find faults in each function; (2) know the internal work to ensure that the internal components work according to specifications. So, in this case there are two types of test cases. First, to demonstrate knowledge of the specific function of the product designed, testing can be conducted to assess whether each function is running as expected. Second, to gain knowledge of how the product works, testing can be done to show how the product works in detail according to its specifications.

There are two kinds of test case approaches, namely white-box and black-box. The white-box approach is a test to show how the product works in detail according to its specifications (Lei & Jiang, 2010) (Jiang, 2012) (Pressman, 2010). The logical path of the application builder software will be tested by providing test cases that will work on a certain set of conditions and loops. Using this method will obtain test cases that ensure that all independent paths in a model have been used at least once. The use of logical decisions on the right and wrong sides, execution of all loops within the constraints and constraints of engineer operations, and use of internal data structures to guarantee its validity. At first glance, it can be concluded that the white box testing approach leads to getting the program 100% correct. The black-box approach is a testing approach to find out whether all software functions have been running well in accordance with the functional requirements set (Jiang, 2012) (Pressman, 2010). This test case aims to show the function of the software that composes the application on how to operate it. This testing technique focuses on the application information domain, namely conducting test cases by partitioning the program input and output domains. The black box method allows the application engineer to derive a set of input conditions that fully utilizes all the functional requirements for a program. This test attempts to find errors in the categories of incorrect or missing functions, interface errors, errors in data structures or external database access, performance errors, and initialization errors and errors.

Application testing is one element of a broader topic often referred to as verification and validation. Verification is a collection of activities that ensure that a software application performs its function. While validation is a collection of various activities that ensure that the application built can meet customer needs. Or in other words, verification is "Are the products we make right?" and validation is “Are we really making the product?”. Validation testing is carried out after all errors are corrected. An indicator of the success of the validation test is if the functions that exist in the software are in accordance with what is expected by the user (Setyaningsih, E., Agustina, P., Anif, S., Ahmad, C., Sofyan, I., Saputra, A., Salleh, W., Shodiq, D., Rahayu, S., & Hidayat, 2022). If the application is made for the customer, acceptance test can be done thus allowing the customer to validate all the requirements. This test is carried out so that customers can find more detailed errors and familiarize customers with understanding the applications that have been made. The form of testing that can be done is alpha and beta testing. Alpha testing is done on the developer side by the customer (Pressman, 2010). The app is deployed in a natural setting with the developer "looking" over the user's shoulder and recording all usage errors and issues. Whereas beta testing is carried out on one or more customers by the end users of the application in a real environment. Developers are usually absent on these tests. The customer records all problems (real or imaginary) encountered during testing and reports to the developer at certain time intervals (Pressman, 2010).

In the end the application product is combined with other system elements and then a series of validation tests are carried out. If the test fails or falls outside the scope of the system development cycle, the steps taken during design and testing can be improved. System testing is a series of different tests with the main objective of working on all elements of the system being developed. Several types of system testing according to Pressman (2010) include recovery testing, security testing, and stress testing.

There are several other aspects in other perspectives that can be used as indicators of a good and optimal test implementation. As stated in the earlier section that the essence of testing is finding software defects and evaluating their quality (Pressman, 2010) (Sommerville, 2011) (Gehring et al., 2017) (Wu, 2010). In terms of quality, it is certainly not easy to justify the quality of an application product or not. The actual level of application product quality is inseparable from how the quality of the test is carried out. Because quality is not a specific concept but an abstract measure, the user can only know and judge that quality is essentially related to the level of service or product and that level is determined from the level of customer satisfaction. Judging from this, it is necessary to set quality standards. Some possible references that can be used to measure the level of quality of application testing are in the form of the quality of the test case itself where application testing can have defects as well and this deficiency can affect the ability of the test to find "bugs". ". The next reference is the quality of the testing process whose stability depends on the test environment. Next is the quality of the test results that can be seen from the test report, as well as the quality of the test clients, namely the report readers. They can immediately feel the effect of the test so that the quality assessment can be considered immediately. Aspects second is the accuracy in choosing the test method and model. Not always a method or model that produces good tests on an application will also be suitable for other applications. The selection of the right test method will certainly contribute to optimal test results. Considerations that can be used in the selection of methods, among others, in terms of time, available manpower, as well as resources and equipment owned. Those three aspects vary in the implementation of the test.

Collaborating several test techniques will certainly increase the reliability of the application being tested because it has passed more than one test case. Application reliability can also be achieved by testing software that implements methods that have been proven to perform well, such as the Bayesian method (Xu et al., 2013) (Cheng et al., 2010) or matrix transformation (Yang et al., 2015) (Yang et al., 2015) (Yang et al., 2015). et al., 2011). The fourth aspect is basing the test on the application architecture. Architectural design provides an overview of the form of the application body that contains components and their relationships. A good understanding of the architecture of an application will be extremely helpful in determining the appropriate test cases and test stages. Architecture-based testing will also assist in deeper flaw detection and prevention. The fifth aspect is that each application test does not need to always create new and special test cases. There is a possibility that the implementation of testing an application is only hosted with other applications. This is possible because the coupler application has actually generated certain actions automatically which could actually behave as test cases for the application under test. If this can be implemented, it will at least reduce the cost of designing new test cases.

The learning application in the form of adaptive learning media developed in this study has received a proper assessment from experts, both learning technology experts, software design experts, linguists, and media users according to what was conveyed by (Wang & Mendori, 2016) regarding the important factors of test the validity and reliability of a product to determine its feasibility. One of the efforts in quality assurance in the development of this application is at the feasibility trial stage carried out three times with the aim of ensuring that media development and the availability of features in the application are in accordance with the needs desired by the user. The results of the first feasibility test managed to capture a lot of suggestions from potential users. After making a number of improvements in accordance with the input suggestions, it is continued with the second trial. A number of suggestions and inputs were also resubmitted by potential users, but the quantity of suggestions submitted was much less than the suggestions in the first feasibility test. After making improvements according to the input suggestions from the second feasibility test, the third trial was then reapplied. The results of this third trial leave suggestions related to the appearance of the application design to improve the adaptive learning application accepted by potential users. In addition, some instruments used in the implementation of this application development research were valid and reliable based on the validity and reliability tests. Finally, it can continue until the stage of testing the effectiveness of the developed media.

Furthermore, the testing process of the effectiveness of adaptive learning applications was conducted using pre-test and post-test designs in the experimental and control groups (Kashani-Vahid et al., 2017). In the experimental group, the learning process was carried out using adaptive learning applications, while in the control group the learning method was used by giving modules and notes. The stage of testing the effectiveness was carried out with a series of pre-test and post-test in the form of ten essay questions sourced from the subject matter. Data. The problem description contains aspects of the HOTS assessment from Facione in (Seventika et al., 2018).

After the analysis, it is known that there is an incredibly significant difference in results where the experimental group using adaptive media in general is better at increasing the average post-test results compared to the module and note user group. Analysis was also conducted on every aspect of critical thinking skills by comparing the gain scores between the experimental and control classes. The increase in the critical HOTS gain score of the experimental group compared to the control group in aspects of Interpretation 93.18 (65.90), Inference 76.74 (44.52), Explanation 96.33 (71.68), Analysis 97.87 (64, 83), Evaluation 63.01 (36.91), and Self-regulation 98.13 (66.34). The overall results show that in the HOTS aspects the experimental group gets a better gain score in the high category, while in the control class it is in the medium category. The results achieved are in accordance with the research submitted by (Nagao & Nagao, 2019); (Drissi & Amirat, 2016); (Bimba et al., 2017); (Tsortanidou et al., 2017) that HOTS can be empowered or improved by providing learning media that are in accordance with the character of the learning style and needs of students.

4. Conclusion

The main goal of application testing is to ensure the product quality of the resulting learning media. There are many parameters that influence to produce quality learning media application products, among others, related to how the environment is during testing, the selection of cases and testing methods, as well as the approach used. Other aspects that contribute to application testing so as to obtain optimal test results include justification in terms of quality from many points of view, accuracy in determining the method and model of the test form, variations in collaborating test techniques, ignoring the form of testing. application architecture, and the possibility of combining (hosting) tests on other applications. In this study, testing steps have been applied carefully following strict testing rules. Based on the analysis of the test results, it can be concluded that adaptive learning media is able to empower students' HOTS with good assessment scores. With the findings from this study, it is expected that all educational application designers always prioritize testing techniques and procedures to ensure the production of quality application products.

5. References

Afandi, A., Sajidan, S., Akhyar, M., & Suryani, N. (2018). Pre-Service Science Teachers’ Perception About High Order Thinking Skills (HOTS) in the 21st Century. *International Journal of Pedagogy and Teacher Education*. https://doi.org/10.20961/ijpte.v2i1.18254

Arief Maulana, Arief Kurniawan, Wini Keumala, Verdian Ramadika Sukma, A. S. (2020). Pengujian Black Box pada Aplikasi Penjualan Berbasis Web Menggunakan Metode Equivalents Partitions (Studi Kasus: PT Arap Store). *Jurnal Teknologi Sistem Informasi*, *3*(1), 23–33. https://doi.org/DOI: http://dx.doi.org/10.32493/jtsi.v3i1.4307

Baoling, W., Xiaohui, C., Peiyi, L., Huilin, J., Hanxiang, G., Jia, L., & Huimin, L. (2020). Development, operational dilemma and tentative idea for construction mechanism of fever clinic in China. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue*. https://doi.org/10.3760/cma.j.cn121430-20200213-00054

Bedjou, K., Azouaou, F., & Berkani, L. (2015). Semantic recommendation of web services in the context of on-line training. *2014 4th International Symposium ISKO-Maghreb: Concepts and Tools for Knowledge Management, ISKO-Maghreb 2014*. https://doi.org/10.1109/ISKO-Maghreb.2014.7033468

Bimba, A. T., Idris, N., Al-Hunaiyyan, A., Mahmud, R. B., & Shuib, N. L. B. M. (2017). Adaptive feedback in computer-based learning environments: a review. *Adaptive Behavior*. https://doi.org/10.1177/1059712317727590

Chang, H. Y., Wang, C. Y., Lee, M. H., Wu, H. K., Liang, J. C., Lee, S. W. Y., Chiou, G. L., Lo, H. C., Lin, J. W., Hsu, C. Y., Wu, Y. T., Chen, S., Hwang, F. K., & Tsai, C. C. (2015). A review of features of technology-supported learning environments based on participants’ perceptions. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2015.06.042

Cheng-Gang, B., J. Chang-Hai, and C. Kai-Yuan. (2010) A reliability improvement predictive approach to software testing with Bayesian method, in IEEE Proceeding of the 29th Chinese Control Conference, July 29-31, Beijing, China, pp. 6031-6036.

Clune, T.L., and R.B. Rood. (2011). Software testing and verification in climate model development, IEEE Journal, Focus: climate change software, September-October, pp. 49-55.

Drissi, S., & Amirat, A. (2016). An adaptive e-learning system based on student’s learning styles: An empirical study. *International Journal of Distance Education Technologies*. https://doi.org/10.4018/IJDET.2016070103

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *34th International Conference on Machine Learning, ICML 2017*.

Jiang, F. and Y. Lu, (2012). Software testing model selection research based on yin-yang testing theory, in

IEEE Proceeding of International Conference on Computer Science and Information Processing (CISP), pp. 590-594

Jin, J., and F. Xue (2011). Rethinking software testing based on software architecture, in IEEE Proceeding of 7th International Conference on Semantics, Knowledge and Grids, pp. 148-151. DOI 10.1109/SKG.2011.32

Jingyun, W., & Takahiko, M. (2015). The Reliability and Validity of Felder-Silverman Index of Learning Styles in Mandarin Version. *Information Engineering Express*.

Kashani-Vahid, L., Afrooz, G. A., Shokoohi-Yekta, M., Kharrazi, K., & Ghobari, B. (2017). Can a creative interpersonal problem solving program improve creative thinking in gifted elementary students? *Thinking Skills and Creativity*. https://doi.org/10.1016/j.tsc.2017.02.011

Khodadi, I., & Abadeh, M. S. (2016). Genetic programming-based feature learning for question answering. *Information Processing and Management*, *52*(2), 340–357. https://doi.org/10.1016/j.ipm.2015.09.001

Kumamoto, H., et.al. (2010). Destructive testing of software systems by model checking, IEEE Journal, pp. 261-266.

Kurniawan, D., Astalini, A., Darmaji, D., Tanti, T., & Maryani, S. (2022). Innovative Learning: Gender Perception of e-Module Linear Equations in Mathematics and Physics. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *4*(2), 92–106.

Kusuma, M. D., Rosidin, U., Abdurrahman, A., & Suyatna, A. (2017). The Development of Higher Order Thinking Skill (Hots) Instrument Assessment In Physics Study. *IOSR Journal of Research & Method in Education (IOSRJRME)*, *07*(01), 26–32. https://doi.org/10.9790/7388-0701052632

Lamas, E., A.V. Dias, and A.M. da Cunha. (2013). Applying testing to enhance software product quality, in IEEE Proceeding of 10th International Conference on Information Technology: New generation, pp. 349-356. DOI 10.1109/ITNG.2013.56

Lei, Y., & Jiang, Y. (2010). Chinese question classification in community question answering. *Proceedings - 2010 IEEE International Conference on Service-Oriented Computing and Applications, SOCA 2010*, 1–6. https://doi.org/10.1109/SOCA.2010.5707167

Lemos, O.A.L., et. al. (2011). "Evaluation studies of software testing research in the Brazilian symposium on software engineering", in IEEE Proceeding of 25th Brazilian Symposium on Software Engineering, pp. 56-65. DOI 10.1109/SBES.2011.30

Martins, A. I., Rosa, A. F., Queirós, A., Silva, A., & Rocha, N. P. (2015). European Portuguese Validation of the System Usability Scale (SUS). *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2015.09.273

Nagao, K., & Nagao, K. (2019). Artificial Intelligence in Education. In *Artificial Intelligence Accelerates Human Learning*. https://doi.org/10.1007/978-981-13-6175-3_1

Nakagawa, E.Y., and J.S. Maldonado. (2011). Contributions and perspectives in architectures of software testing environments, in IEEE Proceeding of 25th Brazilian Symposium on Software Engineering, pp. 66-71. DOI 10.1109/SBES.2011.42

Purmadi, A., & Surjono, H. D. (2016). Pengembangan Bahan Ajar Berbasis Web Berdasarkan Gaya Belajar Siswa untuk Mata Pelajaran Fisika. *Jurnal Inovasi Teknologi Pendidikan*. https://doi.org/10.21831/jitp.v3i2.8285

Reza, H., and S. Lande, 2010, Model based testing using software architecture, in IEEE Proceeding of 7th International Conference on Information Technology, pp. 188-192. DOI 10.1109/ ITNG.2010.122

Sauro, J. (2011). Measuring Usability With The System Usability Scale (SUS). *Measuring Usability*.

Schwan, S., Dutz, S., & Dreger, F. (2018). Multimedia in the wild: Testing the validity of multimedia learning principles in an art exhibition. *Learning and Instruction*, *55*, 148–157. https://doi.org/10.1016/j.learninstruc.2017.10.004

Seechaliao, T. (2017). Instructional Strategies to Support Creativity and Innovation in Education. *Journal of Education and Learning*. https://doi.org/10.5539/jel.v6n4p201

Setyaningsih, E., Agustina, P., Anif, S., Ahmad, C., Sofyan, I., Saputra, A., Salleh, W., Shodiq, D., Rahayu, S., & Hidayat, M. (2022). PBL-STEM Modul Feasibility Test for Preservice Biology Teacher. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *4*(2), 118–127.

Seventika, S. Y., Sukestiyarno, Y. L., & Mariani, S. (2018). Critical thinking analysis based on Facione (2015) - Angelo (1995) logical mathematics material of vocational high school (VHS). *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/983/1/012067

Sommerville, I. (2011). Software engineering, 9th Edition, Pearson Education, USA.

Sugiyono. (2017). MetodePenelitian Kuantitatif, Kualitatif dan R&D. Bandung: PT Alfabet. In *Sugiyono. (2017). MetodePenelitian Kuantitatif, Kualitatif dan R&D. Bandung: PT Alfabet.* https://doi.org/10.1017/CBO9781107415324.004

Sulistyanto, H., Nurkamto, J., Akhyar, M., & Asrowi. (2019). A review of determining the learning style preferences by using computer-based questionnaires on undergraduate students. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1175/1/012209

Tsortanidou, X., Karagiannidis, C., & Koumpis, A. (2017). Adaptive educational hypermedia systems based on learning styles: The case of adaptation rules. *International Journal of Emerging Technologies in Learning*. https://doi.org/10.3991/ijet.v12i05.6967

Vong, S. A., & Kaewurai, W. (2017). Instructional model development to enhance critical thinking and critical thinking teaching ability of trainee students at regional teaching training center in Takeo province, Cambodia. *Kasetsart Journal of Social Sciences*, *38*(1), 88–95. https://doi.org/10.1016/j.kjss.2016.05.002

Wang, J., & Mendori, T. (2016). A Study of the Reliability and Validity of Felder-Soloman Index of Learning Styles in Mandarin Version. *Proceedings - 2015 IIAI 4th International Congress on Advanced Applied Informatics, IIAI-AAI 2015*. https://doi.org/10.1109/IIAI-AAI.2015.284

Wen-hong, L. and W. Xin. (2012). The software quality evaluation method based on software testing, in IEEE Proceeding of International Conference on Computer Science and Service System, pp. 1467-1471. DOI 10.1109/CSSS.2012.369

Wu, H., & Leung, S. O. (2017). Can Likert Scales be Treated as Interval Scales?—A Simulation Study. *Journal of Social Service Research*. https://doi.org/10.1080/01488376.2017.1329775

Wu, Y., Y. Zhang, and M. Lu. (2010). Software reliability accelerated testing method based on mixed testing, IEEE Journal.

Xie, X., Song, W., Liu, L., Du, C., & Wang, H. (2015). Research and implementation of automatic question answering system based on ontology. *The 27th Chinese Control and Decision Conference (2015 CCDC)*, 1366–1370. https://doi.org/10.1109/CCDC.2015.7162131

Xu, X., Cheng, X., Tan, S., Liu, Y., & Shen, H. (2013). Aspect-level opinion mining of online customer reviews. *China Communications*, *10*(3), 25–41. https://doi.org/10.1109/CC.2013.6488828

Yang, M. C., Lee, D. G., Park, S. Y., & Rim, H. C. (2015). Knowledge-based question answering using the semantic embedding space. *Expert Systems with Applications*, *42*(23), 9086–9104. https://doi.org/10.1016/j.eswa.2015.07.009

Yang, Y., L. Lun, and X. Chi. (2011). Research on path generation for software architecture testing matrix transform-based, IEEE Journal, pp. 2483-2486.

Yoshii, A., & Nakajima, T. (2012). Study of a conversational agent system encouraging "real" answers of individuals in a group of acquaintances. *Proceedings - IEEE 9th International Conference on Ubiquitous Intelligence and Computing and IEEE 9th International Conference on Autonomic and Trusted Computing, UIC-ATC 2012*, 143–150. https://doi.org/10.1109/UIC-ATC.2012.124