Development of computerized adaptive testing to measure students' logical thinking skills in science learning

Fitri Wulandari*[1,2], Samsul Hadi[2], Haryanto[2],

[1]Faculty Sains and Technology/ State Islamic University of Sultan Syarif Kasim, Riau

[2]Yogyakarta State University (UNY)

*Corresponding author e-mail: fitri.wulandari@student.uny.ac.id[1], fitriw1974@gmail.com[1]

ABSTRACT

The advancement of information technology had changed conventional methods of testing. Paper based testing and evaluation began to decline due to it takes longer time to process and providing feedback. This study aims to develop a computer-based adaptive test (CAT) to measure students' logical thinking skills in science learning. The CAT development process using the waterfall model covers four main activities, namely: (1) analysis, (2) design, (3) implementation, and (4) testing. The required test material is standardized through a series of trials and item analysis with item response theory (IRT) to obtain item parameters and characteristics which are then used as a database item bank. The procedure for selecting test items uses a fuzzy algorithm using the parameters of item difficulty, item difference power and the response of the participants' answers as input. Based on the results of the system testing, each student receives different test items according to their ability level and the difficulty level of the items received by students according to the characteristics of the item information. Validation Feasibility testing shows the highest grand mean value for student respondents for the use performance aspect was 4.5. This indicated that the result of performance aspects test had a fairly high consistency. The grand mean average value for all aspects, which was above 4, indicated that the development of CAT to measure students' logical thinking skills in science learning is feasible

Keywords:

Computerized Adaptive Testing (CAT), logical thinking skill, Item Response Theory.

Article Received: 18 October 2020, Revised: 3 November 2020, Accepted: 24 December 2020

Introduction

The development of science and technology today demands the ability of individuals to think logically and responsively in decision making and problem solving. The ability to think logically requires mindset skills and cognitive knowledge (Pezzuti et al., 2014) (Seyhan, 2015). The ability to think logically is needed by every individual to be able to solve various complex problems (Sezen & Bülbül, 2011)(Seyhan, 2015), such as the development of skills proclaimed by the Indonesian Ministry of Culture regarding the formulation of the 21st century learning paradigm which is oriented towards students' skills in gathering information, making hypothesis, think logically and collaborate to solve problems (Ministry of Education and Culture, 2013). This also makes one of the goals of science learning, namely to empower students' logical thinking skills(Parmin et al., 2017). However, in fact, based on Trends in the International Mathematics and Science Study (TIMSS) data, student science achievement in Indonesia tends to decline. The results of the 2011 and 2015 TIMSS data analysis in the cognitive realm (knowing, applying, and reasoning) show that the percentage of Indonesian students who answer correctly, especially in the aspect of reasoning abilities, is the part that has not been able to optimally (Martin & Mullis, 2015). PISA (Program for International Student Assessment) Ranking in Science, placed Indonesia in rank 62 out of 70 countries (OECD, 2018). It shown the low level of potential and students' ability in science as such that a breakthrough is needed to improve science skills.

Mathematics and science are subjects that obviously create burden for students since students often have difficulties mastering them (Johnson & May, 2008). Students who have difficulty learning science need attention to improve their abilities. Conceptually, teachers

also find it difficult to provide students with an understanding of science. A teacher, to be successful, must be trained in every necessary dimension: knowledge, abilities, and relationships (Demkanin, 2018). Cognitive abilities can be developed through instruction as well as judgment. Assessment is carried out to determine the extent to which students have successfully received the knowledge provided by the teacher. A good test requires proper assessment construction and management. Paper-based assessment has many weaknesses in its application (Demkanin, 2018). By developing a test-based assessment system that can adapt to students' abilities, it supposed to provide the best solutions in the field of measurement. The development of computer-based systems for examinations to substitute paper-based examination systems has begun to be widely used, for example in English language (Jamieson, 2005;Alwi et al., 2016). Advancement in computer technology really helps computer-based test, since it facilitates administrative processes easely, increase test security, and become more efficient (Khoshsima & Toroujeni, 2017). However, computer-based tests still prone to weaknesses due to it handles large number of test items and unable to provide sufficient information to differentiate student ability scales. Highly performance students get a few easy items so they have little chance of answering it wrong. Likewise, students who have low abilities will get some difficult questions so that they have a small chance to answer correctly. Items like this do not provide sufficient information about the student's level of ability.

Computer-based test for measuring student abilities began to be developed by collaborating adaptive test models in the selection of items, which is called computer adaptive testing (CAT). Computer adaptive testing (CAT) provides a different form of test from computer-based tests that have existed so far. CAT is an adaptive based media where test participants will receive test questions according to their abilities (Thompson & Weiss, 2011). According to Wainer (Howard Wainer, 1990)(H Wainer et al., 2007) adaptive testing is a test which the following questions / items are determined based on the participant's initial answer / response. Computerized adaptive testing able to efficiently shorten the testing time and decrease the number of test items (Cella & Gershon, 2007) it can also accurately estimate the ability of the examinee (Gibbons Robert et al., 2013) and produce the most significant information in measuring the ability of the examinee (Haley et al., 2011). CAT has characteristics, namely (1) provides a question bank containing a collection of test items equipped with statistical characteristics, (2) provides rules for starting the test, whom each test taker does not have to start from the same test item, (3) selects following test items based on the answer/ response to the previous item, if the response is correct then the next test item is more difficult and if the answer is wrong then the next test item is easier. This system is carried out using Item Response Theory (IRT) (Thorpe & Favia, 2012)(Wauters et al., 2010). The test ends when the stopping rule has been reached.

To measure students 'reasoning abilities in science, the test items installed in CAT must be able to measure students' high order thinking (HOT). Rationally modified multiple-choice tests are considered to be a further development of this type of test and aim to measure students' abilities at all cognitive levels, especially higher-order thinking levels. A good test is a test that can accurately measure the test taker's skills, in which the test difficulty index is paired with the test taker's ability. The mechanism of giving is that the item difficulty level will be increased when the test taker answers correctly, and the item difficulty level will be lowered if the test taker answers incorrectly(Haryanto, 2011). In addition, a good test must consider the steps to complete the test item. Innovative testing methods can determine the success of measuring student abilities; thus, it is necessary to develop an instrument to measure logical thinking skills in science using CAT.

Literature Review

This study aims to develop CAT to measure the logical thinking skills of students in grade 9. CAT was developed using the Pressman model (Pressman, 2012), namely the waterfall model which is also called the classic life cycle. This model covers four main activities, namely: (1) analysis, (2) design, (3) implementation, and (4) testing. Initial research was conducted to determine the need for a system that functions as a need assessment. The analysis phase includes a needs analysis and CAT media that are valid, feasible, and efficient to measure students' abilities. The design stage includes CAT development planning and an assessment system using dichotomous data. The implementation phase includes the question bank database building stage and the CAT application, followed by the CAT application testing phase involving IT experts, teachers, and students.

The scope of questions includes science subject under junior high school curriculum. It consists of physics, biology, and chemical materials. The sample of this study was 320 junior high school students from 2 public schools and one private school in Yogyakarta Province, Indonesia. Students were selected using purposive random sampling technique from grade 9 since they supposed to mastered of all science materials.. Student responses are used to calibrate the test by estimating the item parameters.

Item Response Theory (IRT)

Item Response Theory (IRT) is a psychometric theory that provides a basis for measuring the scale of test takers and questions based on the responses given to these questions Hambleton, Swaminatan, & Roger (Hambleton et al., 1991) in (Retnawati, 2014) states that there are three assumptions underlying the item response theory namely unidimensional, local independence, and parameter invariance. Modern testing models with IRT are distinguished based on the number of test item parameters, namely a one-parameter model (Rasch model), two parameters, and three parameters. Van der Linden & Hambleton (Linden et al., 1997), states that these parameters are item difficulty, item difference, and guesswork. The IRT model for the two-parameter dichotomy test items (item difficulty, item difference power) is as follows

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \quad ; \text{ with } i = 1, 2, 3, \ldots, n \quad \ldots.1)$$

Information:

$P_i(\theta)$: the probability of the test taker who has the ability θ to answer item i correctly

θ :level of subject ability (as independent variable)

n : the number of items on the test

e : a natural number whose value is closer to 2.718

a_i: difference power index from item i

b_i : difficulty index from item i

Calibrated items will be stored in the question bank database. The question bank does not only store a set of items whose characteristics are known, but also in the form of a system that organizes the storage of items both from hardware that is ready for use, utilization of items, and deletion of items that are no longer usable. There are several ways to select items from the question bank to be assembled into test kits, one of which is to use the value of the information function.

The item information function is a method for explaining the strength of an item on a set of questions and expressing the strength or contribution of items in revealing the latent ability (latent trait) as measured by the test. Through the item information function, it is known which items match the model so that it helps in selecting test items. Mathematically, the item information function fulfills the following equation

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad ; \quad i = 1,2,3,\ldots n \quad \ldots\ldots\ldots 2)$$

Computerized Adaptive Testing

Computerized Adaptive Testing (CAT) is a method of testing or evaluation using adaptive information technology(Oppl et al., 2017). Adaptive means that giving the next test item

depends on the test taker's response to the previous question so that the test given to each participant can be unique based on the ability level of each participant.

The input to this algorithm is the item difference power, the difficulty level of the items and the response of the test takers' answers. These parameters are processed through a membership function in a fuzzy set. Determination of the first test item for the specialization test is carried out by providing a pre-test (PreTest) to the test takers. PreTest consists of three items with easy, medium and difficult levels. The PreTest results are used as an initial estimate of the test taker's ability and the first test item given will be adjusted to the results of these estimates. The output obtained is the certainty of the test items that have different power and the level of difficulty of the items increases or decreases depending on the response of the test taker's answers. If the answer is correct, the item difficulty level will be increased by 0.2 points, and if the answer is wrong, the item difficulty level will be decreased by 0.2 points. Based on the results of research (Suhardi, 2020) said that the use of stepsize as an alternative strategy for item selection in addition to making items appear more varied, it is also to increase test safety on CAT.

The fuzzy inference model used in this study is Tsukamoto (Yan et al., 1994)(Arya et al., 2014) through four stages, namely (1) Fuzzification, (2) Implication, (3) Inference, and (4) Deffuzification. The formation of fuzzy sets (fuzzification) is the stage for determining the membership value of a variable value. Fuzzy set variables in this research are item difficulty level (b), item difference power (a) and test taker ability (θ). The qualification of the test items was classified based on the difficulty level of the items, namely easy, medium and difficult. Good grain difficulty classification has been set -4 to +4. Besides that, it is also based on the difference in the test items which are classified into low, medium, and high, having a range of 0 to 2. The next stage is Implication, which is the formation of rules, based on the knowledge base (If - Then rule). Next is Inference, which is to determine the extent of the possible area based on the degree of implication function results. The final result will be returned to a crisp value which is called defuzzification using the average center model.

Estimation Ability (Theta)

The method of estimating the ability used in this study is the Maximum Likelihood Estimation (MLE). Assessment of students' abilities is firstly done by calculating the value of pi (θ) and qi (θ) from each test item. In this study, two parameters were used, namely: difference power (ai) and difficulty index (bi), so that the calculation of the value of pi (θ) (the chance of a test taker with the ability character [θ] to answer the test item i-correct) and qi (θ) (the chance that a test taker with ability character [θ] will answer the item in the "i th " test incorrectly) is (Linden et al., 1997):

$$p_i(\theta) = \frac{1}{1+ e^{-Da_i(\theta - b_i)}} \quad \ldots\ldots\ldots\ldots(3)$$

$$p_i(\theta) + q_i(\theta) = 1 \quad \ldots\ldots\ldots\ldots(4)$$

The ability value (θ) is taken in the range -3.00 to 3.0 with steps 0.5. Furthermore, knowing the value of pi (θ), pi (θ), and θ can be calculated the value of Likelihood L (U | θ) with the formula:

$$L(U|\theta) = \prod_{i=1}^{n} {p_i}^n {q_i}^{1-n} \quad \ldots\ldots\ldots\ldots..(5)$$

Information :
n: many test items
u: students' answers to the test items

To find out the character of the students' ability to take the test, it is calculated first using the Maximum Likelihood Estimation L (θ | U) with the formula:

$$L(U|\theta) = \frac{L(U|\theta)}{\sum L(U|\theta)} \quad \ldots\ldots (6)$$

Based on the results of L (U | θ) for the value of θ from -3 to +3, the estimated ability of the test taker is θ from the result of L (U | θ) which is the highest (maximum). The estimated character of

the test taker's ability is determined by the formula:

Estimate θ = Maximum [L(U| θ)] (7)

Results

The CAT web development tool has the following main components: opening page, administration page, teacher page, and student page. The user must enter an email address and password to enter the system (see fig. 2). Each of the user has to be registered by the administrator first.

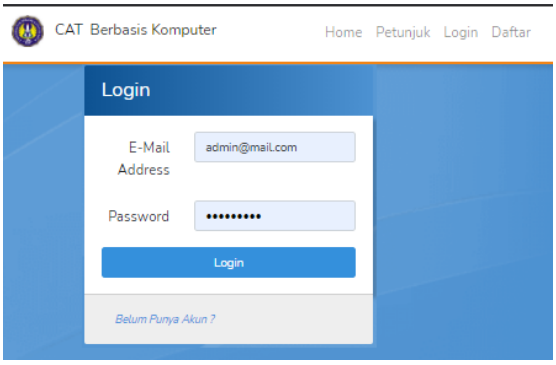

Fig. 1. Login Page

Pretest

The CAT application has two types of tests, namely a preliminary test (PreTest) and an ability test. Preliminary tests are used to initialize students' initial abilities and select the difficulty level of the first item on the ability test. The measurement of students' initial abilities was carried out by presenting three test items with different difficulty levels. The items are selected randomly in one exam which is sorted by level of difficulties namely easy, medium and difficult. If the three items (1,2,3) are answered correctly, the ability level is equal to three (θ = 3), then the items in the difficult category are given. If the items (1,3), (2,3) or (3) are answered correctly then the ability level is equal to 2 (θ = 2), then the items in the difficult category are given. If only items (1,2) or (2) are answered correctly then the ability level is zero (θ = 0), then the items in the medium category are given. If only item 1 or no item is answered correctly then the ability level is equal to -2 (θ = -2), then an item in the easy category is given.

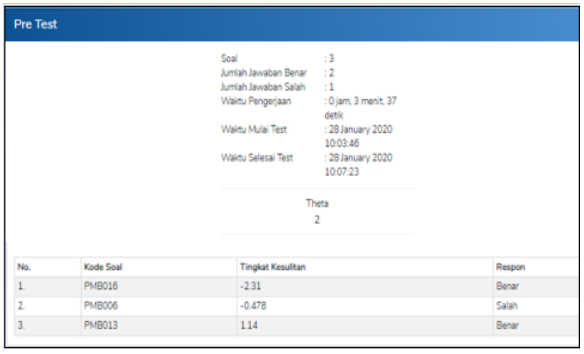

Fig 2. PreTest Results

Figure 2 presents the Pre-test results of one of the test takers. The student responds to item 1 which has a difficulty level of -2.31 with the correct answer, the second item has a difficulty level of -0.478 incorrect answers, and the third item has a difficulty level of 1.34 being answered correctly. Based on the results of the response, the student's initial ability (θ) is high (theta 2).

CAT Trial Results

After knowing the students' initial abilities Furthermore, students will work on Science questions. The first item will be adjusted according to the PreTest results. The form of the question is multiple choice with four answer choices A, B, C, and D, and there is one correct answer. The test page contains information on question numbers, time remaining to work on questions, science questions, alternative answer choices, and answer buttons (see Fig. 3 and 4).

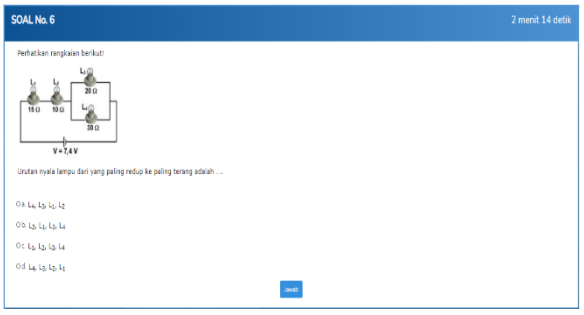

Fig 3. Question Pages

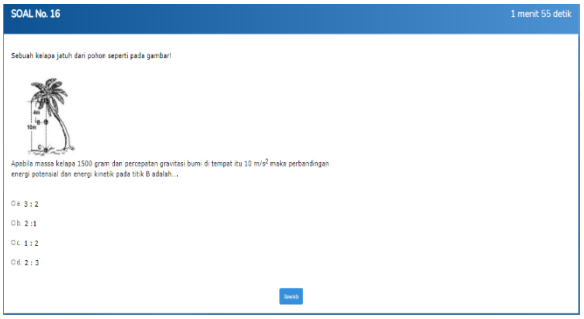

Fig 4. Question Pages

Each question has three minutes to work on. When students start working on the questions, the time will automatically decrease, and if in the time provided, the students have not yet responded to the answers, then the questions will be considered pass and have the wrong answers.

Table 1 shows the answer history of one of the students. If the response is correct then the response is 1 and if the answer is wrong then the response is 0. In the first item, the FIS015 question code has a difficulty level of 1.879, a difference of 0.25, and the answer response is wrong. In accordance with the CAT rules, the difficulty level of the items in the second question will be lowered. Furthermore, using fuzzy inference the second item of question was selected, namely BIO2057, which had a difficulty level of 0.553 and 0.227 of item difference. In this second question, the test participant answered correctly so that for the third item the difficulty level would be increased, and the third item was selected, namely FIS1008 with a difficulty level of 0.694 and a difference of 0.908. And so on until a stop condition is reached. Table 1 also shows that the SEM value shows a decrease, this indicates that the CAT measurement is getting closer to the participants' abilities. After the 13th item, the difference in the SEM value is 0.01 and has reached a stop condition. The test participant answered 17 science questions with 13 items being answered correctly and 4 items being answered incorrectly. These students get a final theta score of 2.5 and if converted to a scale of 100 the score is 91.67.

Table 1. The CAT Result of Science Subject

No	Kode Soal	Tingkat Kesulitan	Daya Beda	Respon	SEM	FIB	Theta
1	FIS1015	1.879	0.25	Salah	74,700	0,0179	-3
2	BIO2057	0.553	0.227	Benar	35,230	0,0806	1
3	FIS1008	0.694	0.908	Benar	27,036	0,1368	3
4	FIS2041	0.913	0.922	Benar	21,159	0,2234	3
5	FIS1011	1.018	0.772	Benar	17,300	0,3341	3
6	FIS2057	1.035	0.899	Salah	0,8796	12,925	2
7	BIO1028	0.47	0.572	Benar	10,696	0,8741	2.5
8	FIS2037	0.772	1.047	Benar	10,063	0,9875	2.5
9	FIS1001	1.202	0.573	Benar	12,161	0,6762	3
10	FIS2053	0.672	1.047	Benar	11,746	0,7249	3
11	FIS1027	1.3	0.383	Benar	11,152	0,8040	3
12	FIS1025	1.34	0.418	Benar	10,571	0,8948	3
13	KIM010	1.658	0.242	Benar	10,347	0,9341	3
14	KIM002	1.864	0.211	Benar	10,180	0,9649	3
15	FIS2060	1.875	0.466	Salah	0,9559	10,944	3
16	BIO2060	0.369	0.71	Salah	0,7498	17,788	2.5
17	BIO1018	-1.054	1.303	Benar	0,7494	17,807	2.5

Based on table 2, it can be seen that the maximum value of L (θ | U) is 0.2861 with the ability position (θ) of 2.5 (figure 5) illustrating that the results of the student's ability test are 2.5. This implies that the chance of students with the ability (θ) = 2.5 to answer the test items correctly is 28%. The graph of ability estimation based on the Likelihood function is shown in Figure 5.

Table 2. Likelihood Estimation Students'1

Ability θ	**L(U|θ)**	**L(θ|U)**
-3	0,0000	0,0000
-2.5	0,0000	0,0000
-2	0,0000	0,0000
-1.5	0,0000	0,0000

-1	0,0000	0,0000
-0.5	0,0000	0,0002
0	0,0000	0,0021
0.5	0,0000	0,0139
1	0,0000	0,0566
1.5	0,0001	0,1445
2	0,0001	0,2427
2.5	0,0001	0,2861
3	0,0001	0.2537
Jumlah	0,0004	1,0000

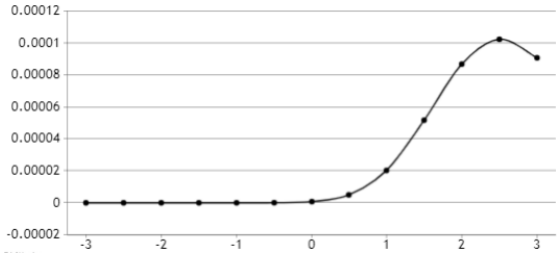

Fig 5. Graph of Student's 1 Likelihood Response Function for Science Subjects

The graph of the ability estimation based on the given items is shown in Figure 6. Even though the 9th to 15th item of theta is always the same, it has not been able to stop the test because the SEM score has not met the specified requirements, namely ≤ 0.01. In the 17th test item the SEM score has reached 0.0001 and the test was stopped with an estimated student ability of 2.5

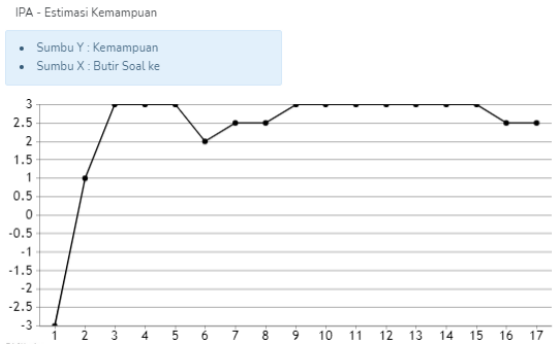

Fig 6. Graph of Estimation of Students' 1 Science Subject Ability

Discussion

The advancement of computer technology has brought a positive impact and created new opportunities for teachers and students. The use of computer technology can be used in the learning process, assessment, measurement and evaluation. This new method helps students to learn many useful scientific concepts more easily. For teachers, this is also used as a medium to measure students' abilities. One of the advantages of computer-based tests is that they can analyze data on a large scale better and more easily (Ahmad et al., 2017). This study aimed to develop an adaptive test that can measure students' ability in logical thinking of science subjects. Accurate or valid tests can produce correct information about a student's skills or abilities (Wynd et al., 2003). Based on expert judgments, the CAT application is considered very good and suitable to be used to measure scientific logical thinking skills. The CAT application able to measure the performance of test takers to complete the test items which were given randomly based on the IRT. Each test taker has answered different questions with different test number. Based on application trials on 73 students, the results showed that students who got a score <60 were 10 or 19.18% of population; students who get a score of ≥60 and <80 are 24 students or 32.88% of population; and students who get a score ≥80 are 35 students or 47.94% of population (Figure 7). Another advantage of CAT is that it can estimate a student's ability level in a shorter time than other test methods (Istiyono et al., 2019). CAT application development can also manage data on a large scale and test results can be seen immediately after students do the test. Accurate test results are very important in obtaining information or data about students' abilities (Andrian et al., 2018). In general, this application can select and provide test items to test takers

based on their abilities, and can measure their abilities accurately.

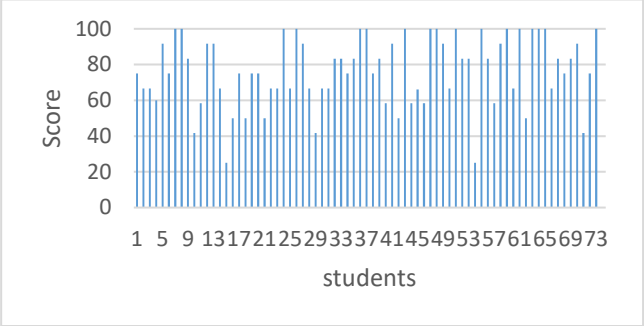

Figure 7. Science Ability Results

CAT validation was carried out to determine the feasibility of the media being developed. Validation Feasibility testing is divided into two parts, namely: (1) alpha testing conducted by the first user (first user / teacher), and (2) beta testing conducted by end users (end users). In the teacher and student user instrument, the response responses use the Likert model with a value range of 1 to 5. The components assessed in the research instrument are in the form of a questionnaire, which includes: (1) user performance aspects, (2) display performance aspects, (3)) the relevance of the test material, and (4) the usefulness aspect. The validation results can be seen in Table 3.

Table 3. Teacher and studens' Responses

Respondents	Aspect			
	Performance	Display	Substance Relevance	Utilization
Teachers	4.4	4.21	4.22	3.97
Students	4.5	4.3	3.9	4.24
Average	4.5	4.26	4.06	4.11

Based on Table 3 above, it shows that respondents have high consistency in validating CAT. The highest grand mean value for student respondents for the use performance aspect was 4.5. This shows that the test results for the performance aspects of use have a fairly high consistency (Retnawati, 2016). The grand mean average value of all aspects, namely above 4, shows that the development of CAT to measure students' logical thinking skills in science learning is feasible to use.

Conclusion

Based on the test results in this study, it can be concluded that the program's ability has succeeded in selecting test items with a difficulty level in accordance with the response of students' answers. Each student receives test items that vary according to their level of ability. This is in accordance with the nature of the CAT theory which demands adaptability in the test. Each student received the correct number of items, and each student received the correct test items according to the characteristics of his ability.

The basis for selecting test items is done by first analyzing, so it is possible that the analysis results are not the same as the type of test items in the question bank database. It is suggested to pay attention to the division of the classification of the level of difficulty of the items and the narrower grain differences in the knowledge base, so that the response to the items that the CAT program raises becomes smoother. It is also necessary to multiply and balance the number of test items for each group. In addition, the types and models of tests that are commonly used in the testing process have many variations, including long answers, complementary tests, matchmaking tests, causal tests, and multiple choice tests. The limitation of this research is that it is not able to handle all types and models of these tests. The type of test developed in this study is limited to multiple choice tests.

Acknowledgement

This study was supported by the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia.

References

[1] Ahmad, Ishak, & et all. (2017). An approach for e-learning data analytics using SOM clustering. *International Journal of Advantaces in Soft Computing and Its Applications*, *7*, 94–112.

[2] Alwi, A., Mehat, M., & Arshad, N. I. (2016). E-Semai Teaching Portal (ESTP): A Preliminary Study in Assisting the Teaching of Bahasa Semai. *International Journal of Advances Soft Computing and Its Application*, *8*(1).

[3] Andrian, D., Kartowagiran, B., & Hadi, S. (2018). The Instrument Development to Evaluate Local Curriculum in Indonesia. *International Journal of Instruction*, *11*(4), 921–934.

[4] Arya, S. H., Abolghasemi, M., Ahmadvand, A. M., & Omran, E. S. (2014). Designing and Implementing the Higher Education Development Fuzzy Expert System in Iran. *Journal of Mathematics and Computer Science*, *8*, 163–179.

[5] Cella, D., & Gershon, Æ. R. (2007). The future of outcomes measurement : item banking , tailored short-forms , and computerized adaptive assessment. *Quality of Life Research*, *16*(1), 133–141. https://doi.org/10.1007/s11136-007-9204-6

[6] Demkanin, P. (2018). Concept formation : Physics teacher and his know-how and know-why. *Journal of Baltic Science Education*, *17*(1), 4–7. https://doi.org/10.4135/9781446288047.n10

[7] Gibbons Robert, D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2013). Development of a Computerized Adaptive Test for Depression. *Arch Gen Psychiatry*, *69*(11), 1104–1112. https://doi.org/10.1001/archgenpsychiatry.2012.14.Development

[8] Haley, S. M., Coster, W. J., & Dumas, H. M. (2011). Accuracy and precision of the Pediatric Evaluation of Disability Inventory computer-adaptive tests (PEDI-CAT). *Developmental Medicine and Child Neurology*, *53*(12), 1100–1106. https://doi.org/10.1111/j.1469-8749.2011.04107.x

[9] Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. In *Sage Publication* (Vol. 21, Issue 2). Sage Publication,Inc. https://doi.org/10.2307/2075521

[10] Haryanto. (2011). Pengembangan Compurized Adaptive Testing (CAT) dengan Algoritma Logika Fuzzy. *Penelitian Dan Evaluasi Pendidikan*, *15*(1), 47–70.

[11] Hole, Y., & Snehal, P. & Bhaskar, M. (2018). Service marketing and quality strategies. Periodicals of engineering and natural sciences,6 (1), 182-196.

[12] Hole, Y., & Snehal, P. & Bhaskar, M. (2019). Porter's five forces model: gives you a competitive advantage. Journal of Advanced Research in Dynamical and Control System, 11 (4), 1436-1448.

[13] Istiyono, E., Dwandaru, W., Setiawan, R., & Megawati, I. (2019). 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.

[14] Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, *25*, 228–242.

[15] Johnson, D., & May, I. M. (2008). The teaching of structural analysis: A report to the Ove Arup Foundation. In *Structural Engineer*.

[16] Khoshsima, H., & Toroujeni, H. (2017). Computer Based Testing : Score Equivalence and Testing Administration Mode Preference in a Comparative Evaluation Study. *International Journal of Emerging*

Technologies in Learning, *12*(10), 35–55.

[17] Linden, V. Der, Wim, J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer Verlag.

[18] Martin, M. O., & Mullis, I. V. S. (2015). TIMSS 2015 International Results in Science. *International Study Center*.

[19] OECD. (2015). *PISA 2015: PISA Results in focus*. https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

[20] Oppl, S., Reisinger, F., Eckmaier, A., & Helm, C. (2017). A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education*, *14*(2), 2–21. https://doi.org/10.1186/s41239-017-0039-0

[21] Parmin, Sajidan, Ashadi, Sutikno, & Fibriana, F. (2017). Performance assessment of practicum work : Measuring the science student teachers' logical thinking abilities. *Man in India*, *97*(13), 141–152. https://doi.org/10.1002/9781444312614.ch8

[22] Pezzuti, L., Artistico, D., Chirumbolo, A., Picone, L., & & Dowd, S. M. (2014). The relevance of logical thinking and cognitive style to everyday prob¬lem solving among older adults. *Learning and Individual Differences*, *36*, 218–223.

[23] Pressman, R. S. (2012). *Rekayasa Perangkat Lunak (Pendekatan Praktisi)* (7th ed.). Andi.

[24] Retnawati, H. (2014). *Teori Respon Butir dan Penerapannya* (1st ed.). Parama Publishing.

[25] Retnawati, H. (2016). *Validitas Reliabilitas & Karakter Butir* (1st ed.). Parama Publishing.

[26] Seyhan, H. (2015). The effects of problem solving applications on the development of science process skills, logical-thinking skills and per¬ception on problem solving ability in the sci¬ence. *Asia-Pacific Forum on Science Learning and Teaching*, *16*(2), 1–31. http://timssandpirls.bc.edu/timss2015/international-results/

[27] Sezen, N., & Bülbül, A. (2011). A scale on logical thinking abilities. *Procedia - Social and Behavioral Sciences*, *15*, 2476–2480. https://doi.org/10.1016/j.sbspro.2011.04.131

[28] Suhardi, I. (2020). Alternative item selection strategies for improving test security in computerized adaptive testing of the algorithm. *Research and Evaluation in Education*, *6*(1), 32–40.

[29] Thompson, N. A., & Weiss, D. J. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assesment Research & Evaluation*, *16*(1), 44–65.

[30] Thorpe, G. L., & Favia, A. (2012). Data Analysis Using Item Response Theory Methodology : An Introduction to Selected Programs and Applications . *Psychology Faculty Scholarship*, 1–33.

[31] Wainer, H, Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.

[32] Wainer, Howard. (1990). *Computerized adaptive testing_ Aprimer*. Lawrence Erlbaum Associates.

[33] Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory : possibilities and challenges. *Journal of Computer Assisted Learning*, 549–562. https://doi.org/10.1111/j.1365-2729.2010.00368.x

[34] Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two Quantitative Approaches for Estimating Content Validity. *Western Journal of Nursing Research*, *25*(5), 508–518. https://doi.org/10.1177/0193945903252998

[35] Yan, J., Ryan, M., & Power, J. (1994). *Using fuzzy logic*. Prentice Hall Inc.

[36] Yogesh Hole et al 2019 J. Phys.: Conf. Ser. 1362 012121