International Journal of Evaluation and Research in Education (IJERE)
Vol. 8, No. 3, September 2019, pp. 376~383
ISSN: 2252-8822, DOI: 10.11591/ijere.v8i3.19642

A computerized adaptive test for measuring the physics critical thinking skills

Aang Zainul Abidin[1], Edi Istiyono[2], Nunung Fadilah[3], Wipsar Sunu Brams Dwandaru[4]
[1,2,3]Department of Physics Education, Graduate School, Yogyakarta State University, Indonesia
[4]Department of Physics Education, Yogyakarta State University, Indonesia

Article Info

Article history:

Received Apr 21, 2019
Revised Jul 30, 2019
Accepted Aug 22, 2019

Keywords:

Adaptive test
Assessment of physics learning
Critical thinking skills
Item response theory
Two-tier multiple-choice

ABSTRACT

Classical assessments that are not comprehensive and do not distinguish students' initial abilities make measurement results far from the actual abilities. This study was conducted to produce a computerized adaptive test for physics critical thinking skills (CAT-PhysCriTS) that met the feasibility criteria. The test was presented for the physics subject of 11th grade high school students with two-tier multiple-choice format. This development research was based on the 4-D model combined with the test development model by Oriondo & Antonio. Eleven experts and 577 11th grade high school students in Kulonprogo, Indonesia, have participated. The media feasibility and the content validity of the items was assessed by experts, while item and abilities parameters were estimated by item response theory. The results obtained: 1) CAT media was declared very feasible and content validity of 136 items was declared valid; 2) all items fitted to partial credit model, the item reliability was classified as good, and the difficulty index of items was good; 3) the results of the CAT-PhysCriTS were equivalent to students academic achievement. Based on the results, CAT-PhysCriTS has fulfilled the requirements as a measuring instrument with measurement times were faster and more comprehensive for large-scale assessments.

Corresponding Author:

Aang Zainul Abidin,
Department of Physics Education, Graduate School,
Yogyakarta State University,
Karang Malang, Catur Tunggal, Depok, Sleman, D.I. Yogyakarta, 55281, Indonesia.
Email: aang.zainul.abidin@gmail.com

1. INTRODUCTION

Critical thinking skills are the main and first things that must be considered in learning in the era of industrial revolution 4.0, as in the learning framework developed by Partnership for 21st Century Learning (P21) [1]. Although the critical thinking is often referred to as the most important element in physics learning, in reality, critical thinking assessment is an area that was often overlooked [2, 3]. Whereas the assessment is extremely needed to find out whether students have had sufficient physics critical thinking skills and what treatment should be given [4]. One reason causing the assessment of critical thinking in physics learning is often overlooked is the lack of availability of tests that specifically measure the physics critical thinking skills. Such problems have actually often been complained of by several researchers [5, 6]. It means that the development of standardized critical thinking tests must be adjusted to the existing needs (physics) so that such problems will not always be repeated.

Another arising problem is the inappropriate test format for measuring critical thinking skills and accommodating material coverage. The forms of tests that are often used are essays and multiple-choices. The essay is the most comprehensive form of a test [7]. However, it has many weaknesses such as a limited measure of high-level thinking, more wasteful in terms of time and cost, high subjectivity, and it is difficult to determine its validity and reliability [8, 9]. Meanwhile, multiple-choice have many advantages such as

easy to implement in large classes, a high level of objectivity, broad material coverage, and can be corrected easily [10]. The disadvantages of the multiple-choice are less comprehensive and the student's thinking process cannot be seen clearly [11]. Based on these strengths and weaknesses, the suitable form of the test must be determined by considering validity, reliability, and comprehensiveness. The solution to this problem is the implementation of two-tier multiple-choice tests [12]. This format consists of several answer options and several options for reasons, so it will require students to think about determining the reasons suitable for their choice of answers [13]. This is in accordance with what was mentioned in [3] that "reason" can change or improve one's mind to act more precisely. In addition, the use of reasons when answering multiple-choice tests is a sensitive and effective way to add value to the comprehensiveness of question items.

The third problem is the use of classical test theory which has many weaknesses also makes measurement results far from the actual abilities. In [14], it was said that in Indonesia, there are still many classical test theories used in analyzing tests and estimating abilities. The classical test theory has many disadvantages, one of which is the characteristics of items depending on the test participants [15]. For example, a question would be an easy question if answered by a clever participant, otherwise, a question would be difficult if answered by the less intelligent participants [16]. Likewise, the score reflecting the participants' abilities depends on the test. If this happens, the actual level of ability of the test participant will never be known [17]. The solution to this problem is the implementation of item response theory (IRT). This is because the level of the participant's ability and estimation of the item parameters is measured accurately [18]. Therefore, in developing the constructs of the physics critical thinking tests must be done with the IRT.

The fourth problem is the efficiency and effectiveness of the test. In recent years, the use of computers in the assessment process began to be implemented by the government of Indonesia in the national high school exam known as a computer-based test (CBT). The national exam uses the same test design for students with the same age or level of education assuming that they have the same ability [19]. In reality, students between schools have significant differences in abilities. Tests that ignore participants' abilities will be suitable for the purpose of measuring the achievement of graduate competency standards because each student has the same competency standards. However, it is less efficient to measure participants' abilities, especially participants with low and high abilities [20]. This is because many question items cannot distinguish test participants [21]. One solution is to apply adaptive tests. The implementation of computer-based adaptive tests is known as computerized adaptive testing (CAT). CAT has the advantages of being more efficient because items that are too easy or too difficult can be avoided so that the length of the test time can be reduced without reducing the measurement precision level, test security is more guaranteed, and the participant's score is immediately known [22, 23]. With the amazing solutions offered by CAT, it would be very useful if applied in measuring the physics critical thinking skills of students.

The relevant research was the development of CAT for assessing problem-solving skills conducted by [24]. The results obtained found that tests developed using CAT were effective to apply. In [13] also developed critical thinking test instruments in the form of two-tier multiple-choice. The results obtained show that two-tier multiple-choice questions were in accordance with the characteristics of physics that prioritize the reasons in answering a problem.

2. RESEARCH METHOD

This is a research and development study (R & D). The resulted product was CAT-PhysCriTS program. This program consists of a website-based CAT media and an item bank concerning physics critical thinking skills. The development model used was the 4-D (define, design, develop, & disseminate) by [25], which was combined with the test development model by [26]. The combination of the two models produced a development procedure shown in Figure 1.

A total of 577 students participated in this study. They were 11^{th}-grade students from six senior high schools in Kulonprogo Regency, Indonesia, in 2018. The subjects during the limited trial stage were 344 students. This number was in accordance with the IRT requirements [27]. Meanwhile, the subjects during the field test stage were 233 students. All participants were selected based on the category of school academic achievement determined by the 2017 National Examination scores. The sample selection was made so that the information obtained can represent students' abilities in general. In this development procedure, data were collected through: 1) questionnaire giving to experts to obtain media feasibility data of CAT and content validity of the items, 2) paper-based test (PBT) in limited trials to find out the characteristics of the items, and 3) CAT in the field trial stage to find out the demographics of students' abilities.

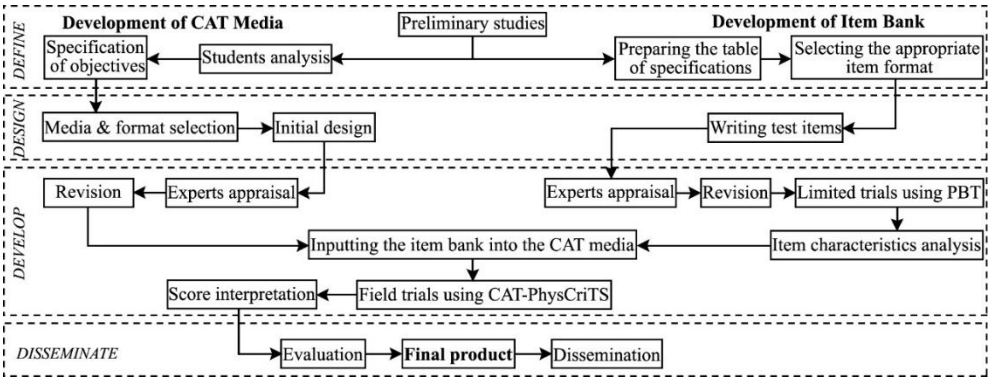

Figure 1. The development procedure

For data analysis, experts' judgment was used to assess the construction of the CAT-PhysCriTS program. The construction was related to media feasibility and content validity of the items. The feasibility of the media was analyzed by comparing the average scores ($\bar{X}$) with ideal average ($X_i = 1/2\,(\max \text{score} + \min \text{score})$) and ideal standard deviations ($sb_i = 1/6\,(\max \text{score} - \min \text{score})$) [28]. If $\bar{X} \leq X_i - 1.8 \times sb_i$ is categorized as "very weak", $X_i - 1.8 \times sb_i < \bar{X} \leq X_i - 0.6 \times sb_i$ is categorized as "weak", $X_i - 0.6 \times sb_i < \bar{X} \leq X_i + 0.6 \times sb_i$ is categorized as "sufficient", $X_i + 0.6 \times sb_i < \bar{X} \leq X_i + 1.8 \times sb_i$ is categorized as "good", and $\bar{X} > X_i + 1.8 \times sb_i$ is categorized as "very good". Meanwhile, the content validity of items was analyzed using the Aikens' formula [29]. The item is said to be valid if it has a coefficient value above the threshold.

The item characteristics were reviewed based on the empirical validity, the test reliability, and the item difficulty index. The empirical validity of the items was determined based on the fittingness between the item with the scoring model. The item fit was based on the results of a limited trial stage. The students' answers were analyzed by a partial credit model (PCM) in the form of polytomous data, namely score 1 if wrong answer and wrong reason, score 2 if correct answer and wrong reason, score 3 if wrong answer and correct reason, and score 4 if the answer and reason are both correct [30, 31]. Scores were analyzed using the Quest program. The results obtained the estimation of the INFIT mean of square value (INFIT MNSQ). Overall, the test is said to be fit if the MNSQ INFIT value is 1.0 ± 0.0 [32]. The goodness of fit of each item can be seen in from the INFIT MNSQ value on each item with the criteria received are 0.77 to 1.30.

The reliability of the scores of test results was estimated according to the IRT based on the reliability of item estimates and the reliability of case estimate [33]. Both estimates were obtained from the analysis of the Quest program. If the reliability value $(R) < 0.67$ is categorized as "poor", $0.67 \leq R < 0.80$ is categorized as "sufficient", $0.80 \leq R \leq 0.90$ is categorized as "good", $0.90 < R \leq 0.94$ is categorized as "very good", and $R > 0.94$ is categorized as "excellent" [34]. The reliability was then proven by a graph of the relationship between information function and SEM obtained from the Parscale program [35]. Meanwhile, the item difficulty index (b) was obtained through analysis of the Quest program. The item was stated to be good if the difficulty index was more than -2.0 or less than +2.0 [36]. If it was close to -2, the item was said to be easy, whereas if it was close to +2 then the item was said to be difficult.

The last data was related to the estimation of the students' ability (θ). Maximum likelihood estimation (MLE) was used as a method of estimating student abilities applied to the CAT-PhysCriTS program [37]. The θ value can be known directly after the student completes the test. Students' abilities were then analyzed by looking at the averages in each school category.

3. RESULTS AND ANALYSIS

3.1. Construction of the CAT-PhysCriTS program

The CAT-PhysCriTS consists of two main components, namely website-based CAT media and an item bank. The construction of the CAT media was related to the format and the algorithm used. The CAT media is a website with a test presentation system as shown in Figure 2, while the example of student test results are shown in Figure 3.

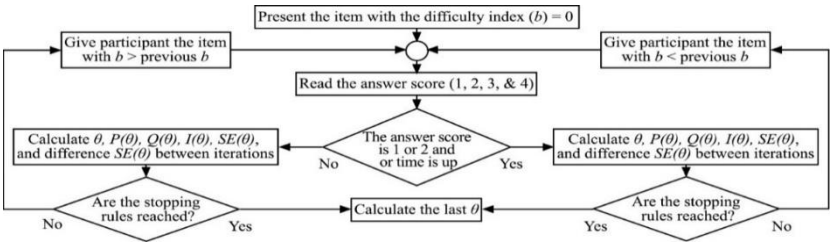

Figure 2. Flowchart of CAT-PhysCriTS

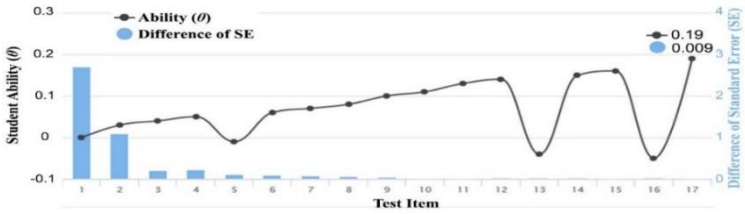

Figure 3. Example of student test results

The CAT algorithm was adapted from the algorithm developed in [38] which was based on IRT, logic, and simple statistics. In Figure 2 we can see that students were given an initial item with moderate difficulty (b = 0) and the initial ability was also assumed to be zero. If students answered correctly (score 3 or 4) then students were given a more difficult question, if wrong (score 1 or 2) then students were given an easier question. When this process took place, the ability (θ) was estimated by the maximum likelihood estimation (MLE). This estimation was done by calculating the ability (θ), the probability of correct answer $\big(P(\theta)\big)$, the probability of wrong answer $\big(Q(\theta)\big)$, information function $\big(I_i(\theta)\big)$, standard error $\big(SE(\theta)\big)$, and difference of *SE* [37]. This algorithm was also illustrated in Figure 3. When students answer correctly difficult item, students' abilities also rise. The decrease in ability recorded in items 5, 13, and 16 was because the student answered incorrectly (score 1 or 2) on the previous item. The final ability of students was 0.19. The test will stop if the items run out or the level of accuracy of the ability estimate (difference of $SE{\sim}0$) has been reached and or if the test time was up.

The CAT media has been assessed for its feasibility by eleven experts. There are two aspects assessed, namely the appearance and function of components. The average score given by the experts in the appearance aspect was 5.82, while the function aspect gets the average score of 8.91. The overall average media feasibility score was 14.73. In accordance with the category in [28], all aspects were categorized very well. Thus, the CAT media has fulfilled the requirements used as a test system in terms of its construction.

The second component of the CAT-PhysCriTS program is the item bank. Specification table of the item bank can be seen in Table 1.

Table 1. Specifications of the item bank

Aspect	Sub-Aspect	Indicator	Physics Subjects					No. of Item
			A	B	C	D	E	
Elementary clarification	Focusing on the question	Formulating a question about events in everyday life.	4	1*	4	4	4	17
	Analyzing an argument	Analyzing statements and determining the similarities or differences in an event given.	4	4	4	4	1*	17
Basic support	Considering source credibility	Examining parts that can be considered to be trusted based on arguments/experiments and their interpretations.	1*	4	1*	4	4	14
	Considering the results of observations	Expressing reasons based on observations on an event.	4	4	4	4	4	20
Inference	Deducting and considering the results of deduction	Interpreting statements and clarifying data.	4	4	4	1*	4	17
	Inducting and considering the results of induction	Making a hypothesis based on several statements/events.	4	4	1*	4	1*	14
Advanced clarification	Defining terms and considering a definition	Defining a statement and considering definitions based on examples in everyday life.	4	4	4	4	4	20
Strategy & tactics	Determining an action	Making decisions based on strategies and solutions, and formulating appropriate alternatives.	4	4	4	1*	4	17
		Number of Item	29	29	26	26	26	136

A = Equilibrium and rotational dynamics, B = Hooke's elasticity and law, C = Static fluid, D = Dynamic fluid, E = Temperature, heat, and heat transfer, and * = Anchor item.

The construction of the item bank is presented in Table 1. The aspects, sub-aspects, and indicators of critical thinking skills used were referring to aspects and sub-aspects developed in [39]. The five physics topics tested were physics materials of the 2013 curriculum (K-13) revised in 2016 for 11th-grade students in the odd semester. The item bank consisted of four test packages with the number of each package was 40 items accompanied by 8 anchor items (total = 136 items). The question format used was two-tier multiple-choice. The first tier contains five answer options and the second tier contains five reason options.

The item bank has been tested for its content validity through the judgment of nine experts (two lecturers and seven physics teachers). Based on the assessment given, then the Aiken's V coefficient was calculated for each item. The results of Aiken's V coefficients of all items were in the range of 0.89 to 1.00. In accordance with what was presented in [29], the tests were said to be valid if they were more than the threshold values in the Aiken's V table. The threshold for 9 raters and approval category of 4 was 0.74. Since Aiken's V coefficients of all items were in the range of 0.89 to 1.00, then all items were stated to have covered the entire area of the contents from the aspect of physics critical thinking skills (valid).

3.2. Items characteristics of the item bank

The items characteristics were reviewed based on the empirical validity, the reliability of scores, and the items difficulty index. The item bank was tested on 344 students. The results of the empirical validity and the reliability values of the test based on analysis using Quest are presented in Table 2.

Table 2. Results of the empirical validity and the test reliability

Parameter	Item Estimation	Testee Estimation
Average value and standard deviation of INFIT MNSQ	1.00±0.05	1.00±0.13
Average value and standard deviation of INFIT t	-0.01±0.49	-0.01±0.85
INFIT MNSQ for each item	0.88-1.18	
Reliability of of item estimate	0.87	
Reliability of case estimate	0.72	

The goodness of fit presented in Table 2 show the overall average value of INFIT MNSQ for item estimation was 1.00 ± 0.05 and for the testee was 1.00 ± 0.13. In addition, the INFIT MNSQ value for each item was in the range of 0.88 to 1.18. In accordance with the explanation in [32], the overall item is said to be fitted with the model if it has an INFIT MNSQ average value of around 1.00 ± 0.00, while each item is fit if it has MNSQ INFIT value of 0.77 to 1.30. Therefore, the result of the goodness of fit as a whole or for each item can be said to be fit with PCM. Thus, the item bank can be declared valid as a measuring instrument.

The reliability of the score in the IRT was indicated by the item separation index and the person separation index as presented in Table 2. The value of the item reliability was classified as good (0.87), while the person reliability was sufficient (0.72). In accordance with explanation in [33], the higher the item reliability, the more convincing the sample matches the tested item, while the higher the person reliability, the more convincing the measurement provides consistent results. Based on these results it can be concluded that the quality of the items was good, while the consistency of students' answers was sufficient.

The value of reliability between one level of ability with other abilities is different. To find out, it can be seen at the curve of the relationship between information function and standard error of measurement (SEM) from the Parscale program presented in Figure 4.

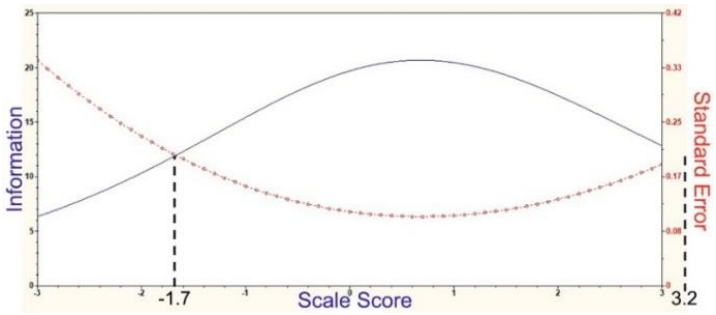

Figure 4. The curve of the relationship beetween information function and SEM

To analyze the curve in Figure 4 is done by looking at the intersection of the line between the information function and the standard error [35]. Based on the curves of the relationship between information function and SEM in Figure 4, it can be concluded that the item bank is suitable or reliable for students with ability values (θ) of -1.7 to 3.2.

The last characteristics of the item bank were related to the item difficulty index (b). The item difficulty index of 136 items was in the range of -0.76 to 0.83. The average of the item difficulty index as a whole was 0.00 ± 0.35, while for each aspect and sub-aspect is presented in Figure 5.

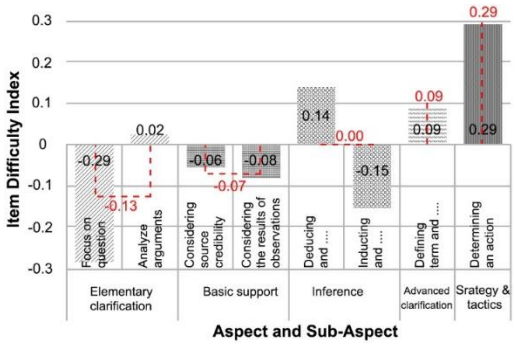

Figure 5. The item difficulty index for each aspect and sub-aspect

The item is said to be good if it had a value of $-2 \leq b \leq 2$ [36]. Because the item difficulty index of the item bank was between -0.76 to 0.83, it can be concluded that the item difficulty index of all items was classified as good. In addition, Figure 5 shows that the sequence of the item difficulty index of each aspect is in accordance with the levels of critical thinking presented in [39] which starting from elementary clarification, basic support, inference, advance clarification, as well as strategy and tactics.

3.3. Test trials using the CAT-PhysCriTS program

After the CAT-PhysCriTS program was tested on 233 students, the lowest and highest abilities (θ) obtained were in the range of 0.03 to 0.60 with an average of 0.15. The physics critical thinking skills obtained based on the school academic achievement categories are presented in Figure 6.

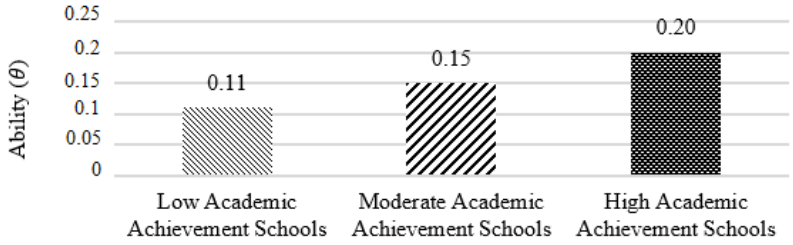

Figure 6. The physics critical thinking skills in each school category

The average ability value shown in Figure 6 is equivalent to the school category. A good measuring instrument can distinguish according to the abilities of students [40]. Based on the suitability between the results of measurement using CAT-PhysCriTS and school academic achievement, we can say the CAT-PhysCriTS program can distinguish the physics critical thinking skills of students.

4. CONCLUSION

The results of the study show that the CAT-PhysCriTS program has met the requirements to be used to measure physics critical thinking skills. Terms that have been fulfilled include: 1) in terms of construction, the CAT media has been declared valid with a very good category in terms of appearance and function aspects. The content validity of 136 items has been declared valid; 2) in terms of the characteristics of the

items, all items were declared fit to the PCM. The quality of the item was good, while the consistency of the answers from the students was sufficient. The item bank was suitable for students with abilities (θ) from -1.7 to 3.2. The difficulty index of all items was classified as good; and 3) the CAT-PhysCriTS program is capable of distinguishing students' critical thinking skills, which is indicated by the measurement results of the CAT-PhysCriTS program equivalent to student academic achievement.

Successfully developed CAT-PhysCriTS will have implications for effectiveness, efficiency, and comprehensiveness in evaluating physics critical thinking skills of students. The CAT-PhysCriTS program presents a faster measurement without reducing its accuracy and is more comprehensive for large-scale assessment. The results of this study can also be used as a reference for education regulators to utilize the CAT system on large-scale assessments such as national examinations. For further research, similar developments can be made for other important skills. In addition, the number of items can be multiplied so that the item difficulty index can be evenly distributed at all levels of student ability.

ACKNOWLEDGEMENTS

We are very grateful to Graduate School, Yogyakarta State University for all the facilities and support in completing this research. We also acknowledge generous support from the government of Indonesia through the Directorate of Research and Community Service, which has supported this project with funding through the graduate school grant research scheme.

REFERENCES

[1] S. E. Saleh, "Critical thinking as a 21st century skill: conceptions, implementation and challenges in the efl classroom," *Eur. J. Foreign Lang. Teach.*, vol. 4(1), pp. 1-16, 2019.
[2] Mundilarto and H. Ismoyo, "Effect of problem-based learning on improvement physics achievement and critical thinking of senior high school student," *J. Balt. Sci. Educ.*, vol. 16(5), pp. 761-779, 2017.
[3] R. H. Ennis, "Critical thinking assessment," *Theory Pract.*, vol. 32(3), pp. 179-186, 1993.
[4] R. W. Elliott, "Understanding faculty engagement in assessment through feedback and dialogues : A mixed methods approach," *Int. J. Eval. Res. Educ.*, vol. 7(3), pp. 167-175, 2018.
[5] F. S. Putri and E. Istiyono, "The development of performance assessment of STEM-based critical thinking skill in the high school physics lessons," *Int. J. Environ. Sci. Educ.*, vol. 12(5), pp. 1269-1281, 2017.
[6] R. S. Damayanti, A. Suyatna, Warsono, and U. Rosidin, "Development of authentic assessment instruments for critical thinking skills in global warming with a scientific approach," *Int. J. Sci. Appl. Sci. Conf. Ser.*, vol. 2(1), pp. 289-299, 2017.
[7] G. N. Khan, N. Ishrat, and A. Q. Khan, "Using item analysis on essay types questions given in summative examination of medical college students : Facility value, discrimination index," *Int. J. Res. Med. Sci.*, vol. 3(1), pp. 178-182, 2015.
[8] K. A. Shaaban, "Assessment of critical thinking skills through reading comprehension," *Int. J. Lang. Stud.*, vol. 8(2), pp. 117-140, 2014.
[9] M. Baig, S. K. Ali, S. Ali, and N. Huda, "Evaluation of multiple choice and short essay question items in basic medical sciences," *Pakistan J. Med. Sci.*, vol. 30(1), pp. 3-6, 2014.
[10] D. D. Kerkman and A. T. Johnson, "Challenging multiple-choice questions to engage critical thinking," *InSight A J. Sch. Teach.*, vol. 9, pp. 92-97, 2014.
[11] E. Istiyono, D. Mardapi, and Suparno, "Development of physics higher order thinking skills test (PysTHOTS) for high school student," *J. Penelit. dan Eval. Pendidik.*, vol. 18(1), pp. 1-12, 2014.
[12] Winarti, Cari, Suparmi, W. Sunarno, and E. Istiyono, "Development of two tier test to assess conceptual understanding in heat and temperature," in *IOP Conf. Series: Journal of Physics: Conf. Series 795 (2017) 012052*, , pp. 1-5, 2017.
[13] F. S. Putri, E. Istiyono, and E. Nurcahyanto, "Development of test instruments for critical thinking skills in the form of reasonable multiple choice (Politomus) in DIY," *Unnes Phys. Educ. J.*, vol. 5(2), pp. 76-84, 2016.
[14] Fajrianthi, W. Hendriani, and B. G. Septarini, "Development of critical thinking test with approach to item response theory," *J. Penelit. dan Eval. Pendidik.*, vol. 20(1), pp. 45-55, 2016.
[15] E. Istiyono, W. B. Dwandaru, and F. Rahayu, "The developing of creative thinking skills test based on modern test theory in physics of senior high schools," *Cakrawala Pendidik.*, vol. 37(2), pp. 190-200, 2018.
[16] A. A. Bichi and R. Talib, "Item response theory : an introduction to latent trait models to test and item development," *Int. J. Eval. Res. Educ.*, vol. 7(2), pp. 142-151, 2018.
[17] O. A. Awopeju and E. R. I. Afolabi, "Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination," *Eur. Sci. J.*, vol. 12(28), pp. 263-284, 2016.
[18] R. K. Hambleton and H. Swaminathan, "Fundamentals of item respon theory,". Newbury Park, Calif: *Sage Publications*, 1991.
[19] H. C. Bagus, "The national exam administration by using computerized adaptive testing (CAT) model," *J. Pendidik. dan Kebud.*, vol. 18(1), pp. 45-53, 2012.

[20] G. Ling, Y. Attali, B. Finn, and E. A. Stone, "Is a computerized adaptive test more motivating than a fixed-item test ?," *Appl. Psychol. Meas.*, vol. 41(7), pp. 495-511, 2017.
[21] E. C. Aybek and R. N. Demirtasli, "Computerized adaptive test (CAT) applications and item response theory models for polytomous items," *Int. J. Res. Educ. Sci.*, vol. 3(2), pp. 475-487, 2017.
[22] M. Rezaie and M. Golshan, "Computer adaptive test (CAT): Advantages and limitations," *Int. J. Educ. Investig.*, vol. 2(5), pp. 128-137, 2015.
[23] H. M. Wu, B. C. Kuo, and S. C. Wang, "Computerized dynamic adaptive tests with immediately individualized feedback for primary school mathematics learning," *Educ. Technol. Soc.*, vol. 20(1), pp. 61-72, 2017.
[24] E. Istiyono, W. S. B. Dwandaru, and R. Faizah, "Mapping of physics problem-solving skills of senior high school students using PhysProSS-CAT," *Res. Eval. Educ.*, vol. 4(2), pp. 144-154, 2018.
[25] S. Thiagarajan, D. S. Semmel, and M. I. Semmel, "Instructional Development for Training Teacher of Exceptional children," Bloomington Indiana: *Indiana University*, 1974.
[26] L. L. Oriondo and E. M. D. Antonio, "Evaluating educational outcomes (Test, measurment and evaluation)," Florentino St: *Rex Printing Company, Inc*, 1998.
[27] E. Muraki and R. D. Bock, "Parscale: IRT item analysis and test scoring for rating scale data," Chicago: *Scientific Softwere International, Inc*, 1998.
[28] S. E. P. Widiyoko, "Techniques for preparing research instruments. Yogyakarta: *Pustaka Pelajar*, 2013.
[29] L. R. Aiken, "Content validity and reliability of single items or questionnaires," *Educ. Psychol. Meas.*, vol. 40(4), pp. 955-959, 1980.
[30] S. Krishnan and N. Idris, "Using PCM to improve the quality of an instrument," *Int. J. Eval. Res. Educ.*, vol. 7(4), pp. 313-316, 2018.
[31] E. Istiyono, "The analysis of senior high school students' physics HOTS in Bantul district measured using PhysReMChoTHOTS," in *American Institute of Physics Conference Proceedings*, vol. 1868, pp. 1-7, 2017
[32] R. Adam and S. Khoo, "Quest: the interactive test analysis system version 2.1," Victoria: *The Australian Council for Educational Research*, 1996.
[33] B. Subali, "Test development with empirical investigation and reliability. Yogyakarta: *UNY Press*, 2016.
[34] B. Sumintono and W. Widhiarso, "Rasch modeling application on educational assessment. Cimahi: *Trim Komunikata*, 2015.
[35] K. Hidayati, Budiyono, and Sugiman, "Using alignment index and polytomous item response theory on statistics essay test," *Eurasian J. Educ. Res.*, vol. 79, pp. 115-132, 2019.
[36] R. K. Hambleton and H. Swaminathan, "Item response theory," Boston, MA: *Kluer Inc*, 1985.
[37] K. T. Han, "Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests," *Appl. Psychol. Meas.*, vol. 40(4), pp. 289-301, 2016.
[38] S. Hadi, "Web-based adaptive test-based computerized development," Yogyakarta: *Aswaja Presindo*, 2013.
[39] R. H. Ennis, "A logical basis for measuring critical thinking skills," *J. Dep. Superv. Curric. Dev.*, vol. 43(2), pp. 45-48, 1985.
[40] E. Istiyono, "IT-based HOTS assessment on physics learning as the 21st century demand at senior high schools: Expectation and reality," in *AIP Conference Proceedings*, vol. 020014, pp. 1-6, 2018.