*International Journal of Instruction* *October 2019 ● Vol.12, No.4*
*e-ISSN: 1308-1470 ● www.e-iji.net* *p-ISSN: 1694-609X*
*pp.OnlineFirst*

Received: 27/11/2018
Revision: 30/05/2019
Accepted: 06/06/2019
OnlineFirst:15/08/2019

# Developing IRT-Based Physics Critical Thinking Skill Test: A CAT to Answer 21st Century Challenge

**Edi Istiyono**
Assoc. Prof., Graduate School, Universitas Negeri Yogyakarta, Indonesia, *edi_istiyono@uny.ac.id*

**Wipsar Sunu Brams Dwandaru**
Dr., Graduate School, Universitas Negeri Yogyakarta, Indonesia, *wipsarian@uny.ac.id*

**Yulita Adelfin Lede**
Graduate School, Universitas Negeri Yogyakarta, Indonesia, *ithaadelfinlede@gmail.com*

**Farida Rahayu**
SMP Negeri 1 Pucang Laban, Tulungagung, Indonesia, *faridarahayu15@gmail.com*

**Amipa Nadapdap**
Graduate School, Universitas Negeri Yogyakarta, Indonesia, *amipanadapdap@gmail.com*

The objective of this study was to develop Physics critical thinking skill test using computerized adaptive test (CAT) based on item response theory (IRT). This research was a development research using 4-D (define, design, develop, and disseminate). The content validity of the items was proven using Aiken's V. The test trial involved 252 students in Yogyakarta, Indonesia. The data were analysed according to partial credit model (PCM). The test reliability was estimated using PCM based on information function (IF) and standard error measurement (SEM), whereas the empirical validity was proven through INFIT MNSQ. The results showed that all items were valid with Aiken's V spread from 0.67 to 1.00 and INFIT MNSQ values from 0.86 to 1.20. The item difficulty index ranges from -0.75 to 1.30, which was a good item difficulty index. The IF and SEM showed that PhysTCriTS was suitable to measure students' critical thinking abilities from -1.80 to 1.50, with the reliability score reaches 0.75. The results of this study have an implication to reduce cheating in a test because each student gets a different item in accordance to the student's ability. The test using CAT may accurately and effectively measure physics critical thinking skill of students.

Keywords: CAT, critical thinking skills, IRT, PCM, Physics, century challenge

**Citation:** Istiyono, E., Dwandaru, W. S. B., Lede, Y. A., Rahayu, F., & Nadapdap, A. (2019). Developing IRT-Based Physics Critical Thinking Skill Test: A CAT to Answer 21st Century Challenge. *International Journal of Instruction*, 12(4), .

## INTRODUCTION

Recently, the 21st century skills which blend with information and communication technology (ICT) become the purposes of global competency (Alismail & McGuire, 2015). Moreover, Živkovic (2016) states that in the 21st century, all professions assume that critical thinking skill is very important to be developed. Therefore, education practitioners are now working on developing students critical thinking skill in order to prepare their graduate for the 21st century competition. The thinking skill can be indicated by students' ability in implementing wise judgment or producing reasonable critique (Husamah, *et. al*, 2018). The ability increases the performance of graduates to cooperate successfully, think analytically, and solve problems efficiently in the workplace. That intended ability is called higher order thinking ability.

Critical thinking belongs to Higher Order Thinking Skill (HOTS) (Ennis & Weir, 1985). Cottrell (2011) states that there are six sub-categories of HOTS, three of which are sub-categories of analysis, evaluation, and creating. Those sub-categories underlie the critical in dealing with a problem, therefore, HOTS are useful in the global competition.

Ibrahim (2007) defines critical thinking as a kind of thinking which systematically investigates one's thinking process using evidence and logic. In line with Ibrahim, *et. al* (1985) states that critical thinking is rational reflective thinking focused on deciding what to believe or do. Reflective thinking requires a systematic evaluation of high-standard thinking patterns including skills to evaluate information and obtain the correct and logic solutions. By using critical thinking, one can make the right decision by considering the systematic and analytic evaluation. Related to critical thinking, Fischer (2009) mentions several abilities, namely: (1) recognizing the problem; (2) finding ways that can be used to solve problems; (3) collecting and compiling necessary information; (4) understanding and using appropriate language, analyzing data, assessing facts, and evaluating statements; (5) recognizing a logical relationship between problems; (6) drawing the necessary conclusions and similarities; (7) examining the similarities and conclusions. Those capabilities are identified as critical thinking skills. As a conclusion, critical thinking is what instructors should prioritize and promote as a practice skill for students (Thaneeranon *et al.*, 2016).

In order to optimize the critical thinking skill, educational practitioners need to formulate a suitable assessment to monitor students' ability. Many in-class assessments have been applied throughout the history of education, but testing seems to be the most favorite method. The discussion about assessment leads to the assessment design to assess the core competencies for special skills, such as HOTS. In the history of testing, multiple choice still becomes the favorite type of test (Arif, 2014). Common multiple-choice test is done by judging the answer, which is called dichotomy scoring system. Dichotomy scoring system is also known as classical test theory (CTT) that has either true or false answer (Hambleton & Swaminathan, 1991). However, multiple choice cannot be used to measure higher order thinking skills, so it has to be modified (Brookhart, 2011). Assessments in the form of multiple choices should be modified since it leads students to think of choosing the inappropriate answers. An alternative modification of multiple-choice test is reasoning multiple-choice (Istiyono, *et. al*, 2014).

This type of test is able to be used to test students' higher order thinking skill. The other alternative is an essay test (Brookhart, 2010). However, the essay-type test is difficult to apply in a large scale.

As a teacher, designing a large-scale assessment such as summative assessment in one tenure requires an enormous effort. Even Indonesian government is still struggling in implementing an appropriate assessment system for the large-scale assessments within the country. In developing a test, it is necessary to prepare a blueprint in the form of a table including the activities and the competencies tested (Al-Fallay, 2018). In Indonesia, the government is now starting to use ICT as a large-scale assessment medium to replace paper and pencil.

To be in synchronicity with the development of the century, computer-assisted tests have been widely applied for large-scale tests. The study of the use of computers as a test medium begins in the early 1970s. Tests using computers have many advantages. Besides, it is able to produce the same test with low cost, and also able to minimize human errors because the scoring is done by the computer. Recently, computer-assisted tests is involved as an innovative medium for testing. Computer-assisted tests can be applied to various types of test, one of which is adaptive testing (Hosseini *et. al*, 2017).

CAT is said to be the most important psychological assessment development. By using CAT, the difficulty index of the next items will change depending on the previous question answered (Finkelman *et. al*, 2014). Hadi & Haryanto (2012) explains several principles of CAT, including: (1) at first, the test takers are given test items with standard or moderate difficulty index, items with difficulty index close to zero, (2) if the test takers answer correctly (on a scale of 3 and 4), the test takers will then get a higher difficulty index item, (3) if the test takers answer wrongly (on a scale of 1 and 2) then they will get item that has lower difficulty index. The CAT software created uses an algorithm system to display the items in accordance with the student's abilities. The algorithm is applied to make a decision concerning the next item to be given, corresponding to the student's answer of the previous question. The next items are determined based on the IRT theory, logics, and simple statistics. This method can motivate the students to show their maximum abilities, therefore, the use of CAT is assumed to be able to motivate the students to solve problems.

IRT has sub-scoring systems called nominal response model (NRM), rating scale model (RSM), graded response model (GRM), PCM, and generalized partial credit model (GPCM). Unlike CTT, IRT has various scores (polytomy), which is not only fully right or wrong. PCM in IRT can be called as a stepwise solution of polytomy scored items; the item parameters are interpreted as the difficulties of the steps (Verhelst & Verstralen, 2008). The purpose of using PCM is to get a better score as the ability is better (Widhiarso, 2010). PCM can be used to interpret the science and critical thinking ability (Istiyono *et. al*, 2014). PCM is considered to be a scoring system that teacher needs to measure the students' ability, which is not only limited to score the students' response. There are at least two strengths offered by PCM, the simplicity to implement model's formulation in practice and the availability of the PCM in a range of software packages (Linden & Humbleton, 1997).

According to Linden and Humbleton (1997), PCM is designed to analyze a wide range of abilities on the basis of their level of response. However, the mapping of response categories is a little less straightforward than for right-or-wrong scoring. In line with the previous statement, Bond & Fox (2015) agree that PCM considers the possibility of tests in which one or more intermediate level of success might exist between a complete failure and a complete success. PCM which is developed according to Rasch model scoring system has different difficulty level for every item, depending on every person's trait. In other words, PCM contains two sets of parameters: one for the person and one for items (Linden & Humbleton, 1997). Muraki & Bock (1997) have mentioned the formula of PCM given in Eq. (1).

$$P_{ig}(\theta) = \frac{\exp\left[\sum_{g=0}^{l}(\theta - b_{ig})\right]}{\sum_{h=0}^{m}\exp\left[\sum_{g=0}^{h}(\theta - b_{ig})\right]} \quad g = 1,2,3,\ldots,m+1, \tag{1}$$

with $P_{ig}(\theta)$ is the probability of student with ability $\theta$ to answer $i$ item correctly, $\theta$ is the student's ability, $m$+1 is the amount of $i$ item category, and $b_{ig}$ is the threshold index of $i$ item category. The formula can be modified according to the items, for example, items with score 1, 2, 3, and 4. The differences in the score represent the different probability of each student. The formula is given as follows:

$$\sum_{g=0}^{0}(\theta - b_{ig}) \equiv 0 \quad \sum_{g=0}^{k}(\theta - b_{ig}) \equiv \sum_{g=1}^{k}(\theta - b_{ig}) \tag{2}$$

Linacre (2006) has stated that $b_{ig}$ is also interpreted as a node when two different categories have the same probability to be chosen on the different trait level.

In case of physics learning, Hadi & Handhika (2015) state that physics learning at school should be conducted using a scientific approach in order to be more meaningful and able shape the students' characters. The scientific approach in physics learning is able to be used to develop students' soft skills and promote their abilities to think creatively and critically. Istiyono (2017) states that there are many teachers who have failed to give questions regarding students' thinking skill; they tend to give questions that measure student's memory (lower order thinking skill-LOTS). Both concepts of scientific approach in learning and critical thinking, are in accordance with the concept of Curriculum 2013, the curriculum impelemented in Indonesia.

According to the previous explanation related to the 21$^{st}$ demand on education and its importance, also the requirement of the physics curriculum in Indonesia, The use of CAT to facilitate teachers in assessing students' critical thinking skill is worth to develop. Through this reasearch, sets of tests that are reliable and valid administrated using CAT are prepared.

## METHOD

### Development Model

This is a development research using Thiagarajan flow, which is a 4-D development model consisting of 4 stages; define, design, develop, and disseminate. (1) Define stage is a defining phase that aims to collect information related to the development carried out. The information obtained might come from literature studies or preliminary studies. The preliminary study aims to determine the specification of instruments test developed. (2) The design stage is a step to design the product developed. In this phase, the blue print, items, and test score guidelines are written. (3) The development phase consists of two activities. The first is validation activities to assess the product feasibility, which are carried out by experts in their fields, and the suggestions given are used to improve the product within the research. The second activity is empirical testing or product testing on real target subjects. After it is proven that PhysTCriTS is feasible to measure physics critical thinking skill, assembling of PhysTCriTS is conducted into CAT. (4) The dissemination stage is the final stage of product development such that the product can be used by others.

### Participants and Research Samples

In the validation process, two physicists, two experts (lecturers) on assessment and one practitioner were involved to assess the test according to the physics materials tested, the language used, and the test construction. For the trial test, 252 high school students were involved. Table 1 shows the distribution of the participants.

Table 1
The Distribution of Test Trial Participants

| Schools | Package | | Participants |
|---|---|---|---|
| | A | B | |
| SMA N 3 Bantul | 43 | 44 | 87 |
| SMA N 2 Bantul | 39 | 41 | 80 |
| SMA N 1 Bantul | 41 | 44 | 85 |
| Total | 123 | 129 | 252 |

### Data Collection and Analysis

The data collection technique used was PhysTCriTS. PhysTCriTS is a test developed to measure critical thinking skill of the $10^{th}$ grade students of Senior High School. The instruments used were two sets of tests (package A and B) and a questionnaire to validate the test. First, both sets of tests were validated by the experts using a questionnaire. The questionnaire consisted of four interval scales in order to assess the properness of the tests.

Product validation is carried out in two stages: content validity and empirical validity. Content validity used the Delphi method. The validation process involves physics education experts, measurement experts and media experts. The results of experts' validation were analyzed using Aiken's formula. The Aiken formula used is as follows:

$$V = \frac{s}{n(c-1)} \tag{3}$$

with *r* is the experts' remarks, *n* is the number of points, *c* is the biggest scale for evaluation, and $I_0$ is the smallest scale of evaluation, that is:

$$s = r - I_o. \tag{4}$$

The validation media questionnaire data was analyzed and converted into several value of intervals therefore the criteria were obtained as in Table 2.

Table 2
Feasibility Categories

| No | Score | Category |
|---|---|---|
| 1 | More than M + 1,8 SD | Very feasible |
| 2 | M + 0,6 SD to M +1,8 SD | Feasible |
| 3 | M – 0,6 SD to M + 0,6 SD | Fair |
| 4 | M – 1,6 SD to M – 0,6 SD | Less feasible |
| 5 | Less than M – 1,8 SD | Not feasible |

After fulfilling the validity content with results and declared for its feasible or very feasible then empirical validity is carried out through two stages; limited trials and wide-scale trials. The test trial was conducted to test the empiric validation of PhysTCriTS. Students' response on the test were scored using polytomi category, i.e., 1; 2; 3; and 4. The data collected were analyzed using partial credit model (PCM) for testing the item-fit. PCM is the development of the Rasch Model, which is, a 1-PL model. The data analysis was performed on several aspects, including (1) the goodness of fit, (2) reliability, (3) item difficulty index, and (4) information function and standard error measurement (SEM).

## FINDINGS

### Content Validity

The PhysTCriTS content validity was measured according to the material, construction, and the language used. Two physicists, two assessment lecturers, and one practitioner conducted the expert judgments. The results of the Aiken's for critical thinking skill tests start from 0.67 to 1.00, respectively.

### Empirical Validation

The result of the empirical validation is determined by the suitability (goodness of fit) of the instrument item. The fitness of the test items is determined by observing the average value of INFIT MNSQ, besides, standard deviation can also be considered. If the INFIT MNSQ average is around 1.00 and the standard deviation is 0.00 or the INFIT t rate is close to 0.00 and the standard deviation is 1.00, the whole test fits the PCM. Moreover, the item acceptance limit used is 0.6 to 1.21 for INFIT MNSQ values. The results show

that the overall goodness of fit scores (INFIT MNSQ) of critical thinking skills tests in PhysTCriTS is from 0.86 to 1.20, respectively (see Figure 1).

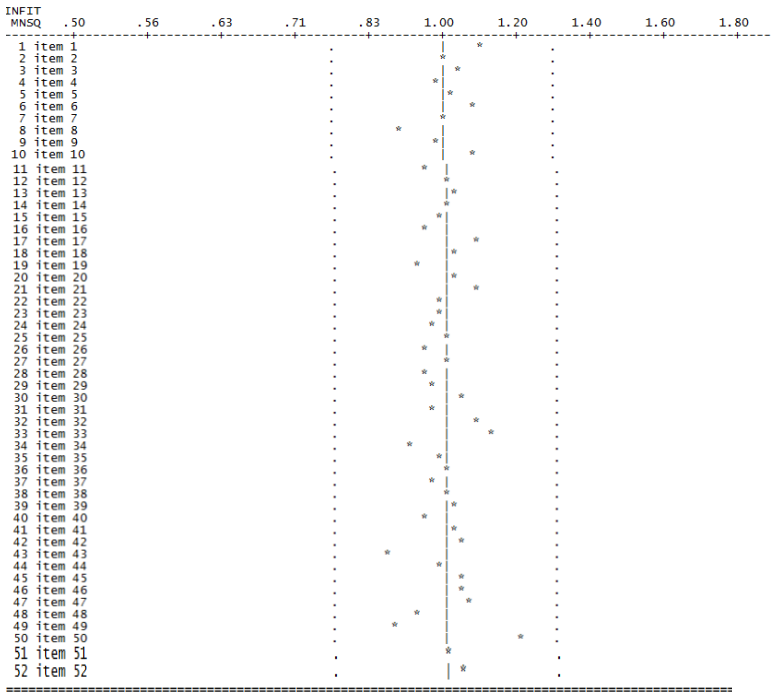



Figure 1
The INFIT MNSQ of PhysTCriTS

Figure 1 shows the item spread in the form of points spread. The figure shows that all items are fit (valid) as they are at values around 0.86 to 1.2. Validity assumptions on the Rasch model refer to INFIT MNQ with value ranges from 0.6 to 1.21; OUTFIT MNSQ with values of 0.11 to 1.17 (Huang *et. al*, 2018). This means that the items are feasible to be used to measure the students' ability in physics learning according to the given subject material.

**Item Difficulty Index**

The calculation of the difficulty index for 52 items concerning critical thinking skills in two packages shows that the difficulty indexes are in the range of -0.75 to 1.30, respectively. The index of difficulty for each aspect can be observed in Figure 2.

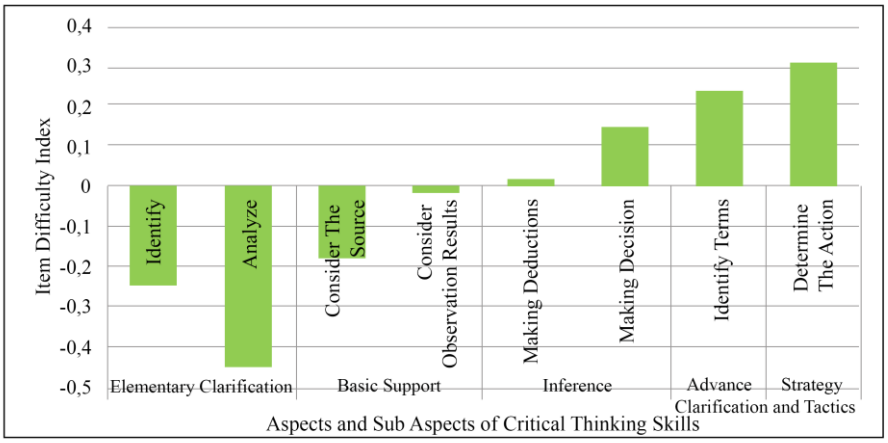



Figure 2
Item Difficulty Index of each Aspect and Sub-aspect of PhysTCriTS

**Information Function and SEM and Test Reliability**

The information function and SEM are inversely related. The result of the information function and SEM on PhysTCriTS are sequentially presented in Figure 3. Figure 3 shows that the information function and SEM is spread from -1.80 to 1.50, respectively.

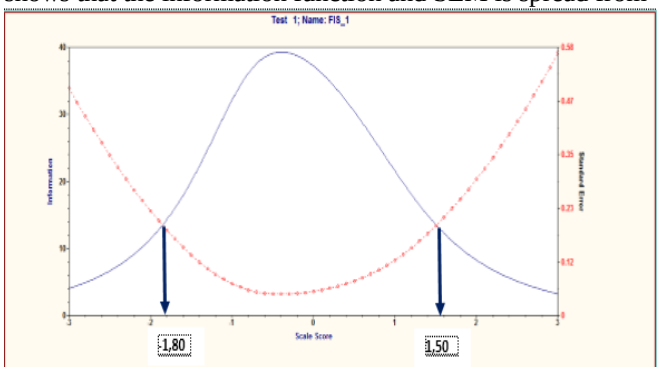



Figure 3
Information Function and SEM of PhysTCriTS

Figure 3 shows that the information function and SEM is spread from -1.80 to 1.50, respectively. The SEM is marked by the red curve (dashed-line), whereas the total information function is shown by the blue curve (solid-line). The interval between the secant of the two curves is the ideal ability border in answering test items, which is in this case is from -1.80 to 1.50. The item reliability value of critical thinking skills test obtained from the summary value of item estimate is 0.75. The reliability coefficient which is considered to be appropriate for measurement is > 0.70 (Lima *et. al*, 2018).

**Assembling PhysTCriTS into CAT**

The developed critical thinking skills test instrument, which is valid and reliable according to the analysis was then imported into CAT system. CAT is able to present the problems (questions) to the students in accordance with their ability. Figure 4 shows the example of a PhysTCriTS screen view.

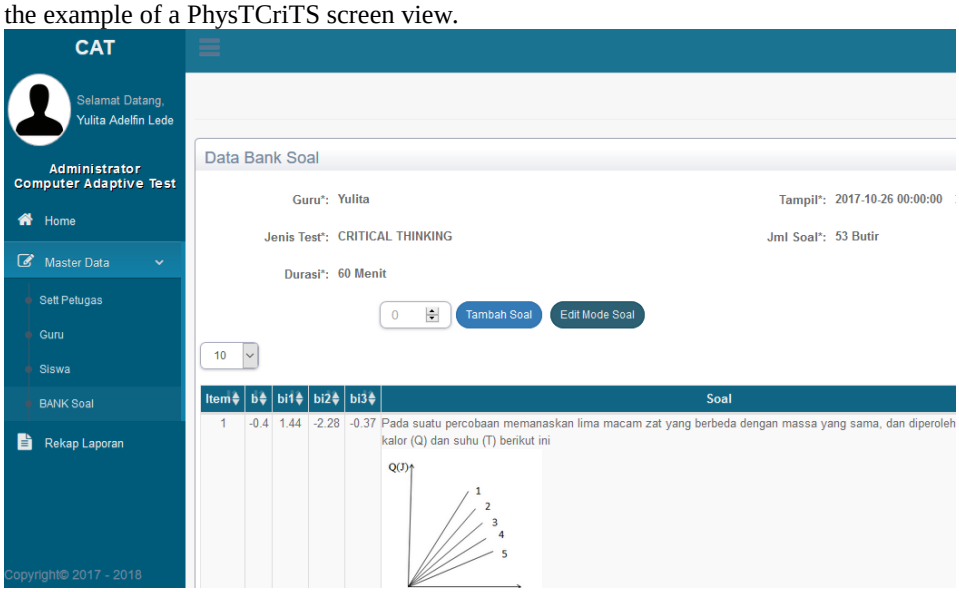



Figure 4
An Example of Installing of PhysTCriTS's Items into CAT

**DISCUSSION**

In order to provide ready-to-use sets of tests, the tests have to go through several analysis, i.e., validity and reliability analysis. The content validity test of PhysTCriTS is determined through Aiken V's formula. The result shows that the content validity of PhysTCriTS starts from 0.67 to 1.00, respectively. The content validity based on Aiken's V is said to be good with a value of more than 0.80 (Azwar, 2017). Therefore, the PhysTCriTS test items are considered to be valid.

Following the content validity test, empirical validity analysis was performed to find out the items' validity based on the test trial administrated to 252 students. The result of the empirical validation is determined by the suitability (goodness of fit) of the instrument item. Being fit means that the actual items are close enough to the Rasch Model's requirements to be counted as linear interval scale measures (Bond and Fox, 2015). The empirical validity test results (See Figure 1) show that the overall goodness of fit scores of critical thinking skills tests in PhysTCriTS is in the range of 0.86 to 1.20, respectively. The validity assumption on the Rasch model refer to INFIT MNSC with values that ranges from 0.6 to 1.21; OUTFIT MNSQ with values of 0.11 – 1.17 (Huang *et. al*, 2018). Besides that, the average score of INFIT MNSQ is 1.00 and its standard

deviation is 0.00, therefore PhysTCriTS is considered to fit the model. According to the result of the analysis, it can be concluded that PhysTCriTS which consists of 52 items (questions) concerning critical thinking skills is considered to be valid because it fits the partial credit model (PCM).

Completing the validity analysis, information function, SEM and reliability test were performed. The information function and SEM are inversely related. This is in accordance with the theory in Hambleton & Swaminathan (1991) which states that the SEM and the information function are inversely proportional, the higher the SEM, the lower the information function becomes, and vice versa. This relationship shows that the ability of students is in accordance with the developed instrument. Based on Figure 3, it can be seen that the critical thinking skills test is suitable for students with the ability of $-1.80 \leq \theta \leq 1.50$, so students can take the critical thinking skills test with medium to high abilities as well.

The item reliability value of critical thinking skills test obtained from the summary value of item estimate is 0.75. The test reliability based on the summary value of case estimate is 0.74. Based on the result, PhysTCriTS is considered to be valid. The conclusion is in accordance with the interpretation of reliability determined. If the test reliability value is in the range of 0.67 to 0.80, then the test reliability is considered to be quite good.

In order to be able to be assembled into CAT, the item difficulty index need to be determined. The item difficulty index is the degree of difficulty per step. The difficulty index determines the next items that students will face after answering a specific item. If the test takers answer the item correctly (on a scale of 3 and 4), they will then face the more difficult item, but if the test takers answer the item wrongly (on a scale of 1 and 2) then they will get easier item (Hadi, 2012). According to the item difficulty index analysis, 52 items concerning critical thinking skills in two packages of PhysTCriTS show that their difficulty indexes are in the range of -0.75 to 1.30, respectively. These results show that the difficulty indexes of these items are good. The difficulty index of the test item, which is categorized as good is in the value range of -2 to 2 (Retnawati, 2014; Mardapi, 2017).

The results of the analysis toward the quality of PhysTCriTS show that PhysTCriTS is considered to be valid, reliable, and has a good range of difficulty which is able to cover all levels of students' ability. Considering those conclusions, PhysTCriTS is ready to be assembled into CAT. CAT is based on item response theory, selects items from an item bank that are most appropriate for each child, thereby minimizing the number of items needed to ensure an accurate score (Huang, *et.al*, 2018: 2). This is also supported by the study of Whiley *et. al* (2017) which states that the development of students' critical thinking skill is affected by a good study environment management, e.g.: concerning the curriculum design and learning evaluation. Learning evaluation that begins with a good evaluation using CAT may develop students' critical thinking skill. Figure 5 shows the example of PhysTCriTS item appearance in CAT.

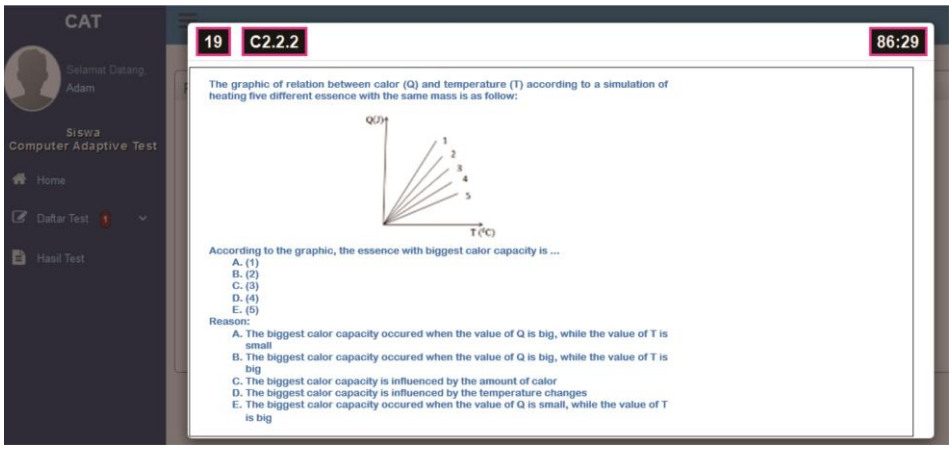
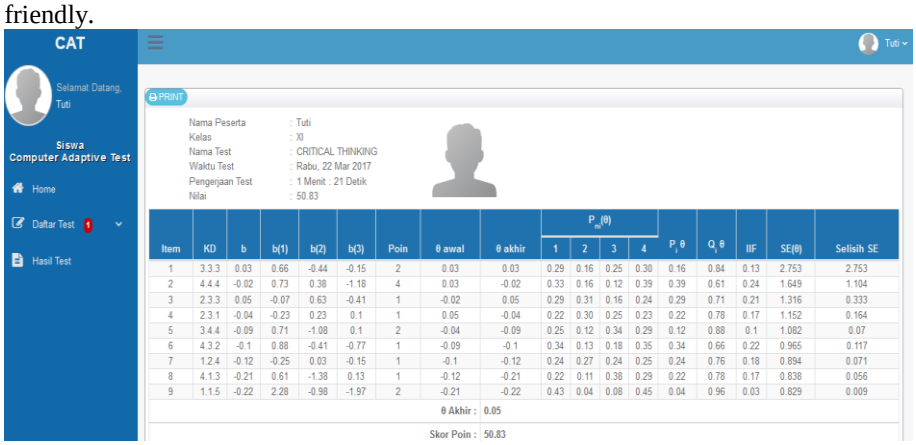



Figure 5
The Example of PhysTCriTS Item Appearance in CAT

PhysTCriTS which implements CAT system is able to present the problems (questions) to the students in accordance with their ability as Finkelman et al, (2014) state that the next item will then be different, the difficulty index of the next items will change depending on the previous question answered. In addition, CAT is able to provide quick feedback for both student and teacher. Figure 6 shows the feedback provided by CAT. CAT is said to be the most important psychological assessment development since it is able to minimize human error in administrating the test, and also to make scoring cost-friendly.




| Item | KD | b | b(1) | b(2) | b(3) | Poin | θ awal | θ akhir | $P_{ni}(\theta)$ 1 | 2 | 3 | 4 | $P_i\theta$ | $Q_i\theta$ | IIF | SE(θ) | Selisih SE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3.3.3 | 0.03 | 0.66 | -0.44 | -0.15 | 2 | 0.03 | 0.03 | 0.29 | 0.16 | 0.25 | 0.30 | 0.16 | 0.84 | 0.13 | 2.753 | 2.753 |
| 2 | 4.4.4 | -0.02 | 0.73 | 0.38 | -1.18 | 4 | 0.03 | -0.02 | 0.33 | 0.16 | 0.12 | 0.39 | 0.39 | 0.61 | 0.24 | 1.649 | 1.104 |
| 3 | 2.3.3 | 0.05 | -0.07 | 0.63 | -0.41 | 1 | -0.02 | 0.05 | 0.29 | 0.31 | 0.16 | 0.24 | 0.29 | 0.71 | 0.21 | 1.316 | 0.333 |
| 4 | 2.3.1 | -0.04 | -0.23 | 0.23 | 0.1 | 1 | 0.05 | -0.04 | 0.22 | 0.30 | 0.25 | 0.23 | 0.22 | 0.78 | 0.17 | 1.152 | 0.164 |
| 5 | 3.4.4 | -0.09 | 0.71 | -1.08 | 0.1 | 2 | -0.04 | -0.09 | 0.25 | 0.12 | 0.34 | 0.29 | 0.12 | 0.88 | 0.1 | 1.082 | 0.07 |
| 6 | 4.3.2 | -0.1 | 0.88 | -0.41 | -0.77 | 1 | -0.09 | -0.1 | 0.34 | 0.13 | 0.18 | 0.35 | 0.34 | 0.66 | 0.22 | 0.965 | 0.117 |
| 7 | 1.2.4 | -0.12 | -0.25 | 0.03 | -0.15 | 1 | -0.1 | -0.12 | 0.24 | 0.27 | 0.24 | 0.25 | 0.24 | 0.76 | 0.18 | 0.894 | 0.071 |
| 8 | 4.1.3 | -0.21 | 0.61 | -1.38 | 0.13 | 1 | -0.12 | -0.21 | 0.22 | 0.11 | 0.38 | 0.29 | 0.22 | 0.78 | 0.17 | 0.838 | 0.056 |
| 9 | 1.1.5 | -0.22 | 2.28 | -0.98 | -1.97 | 2 | -0.21 | -0.22 | 0.43 | 0.04 | 0.08 | 0.45 | 0.04 | 0.96 | 0.03 | 0.829 | 0.009 |

Figure 6
The Feedback Appearance Provided by CAT

Figure 6 shows the output results of PhysTCriTS using CAT of one of the students. The figure contains the student's identity, item identity which is conducted, and the student's

physics critical thinking ability (θ) which is stated in logit scale based on IRT according to PCM and in the scale of 100.

The results of this study have implications for reducing cheating (fraud) in the thest because each student gets a different item according to the ability of the students. Furthermore, the test with CAT can also accurately measure students' critical thinking skills with more effectiv and efficient time and energy. This is also reinforced by the results of Huang, et. al (2018) which states that CAT based on IRT, selects items from an item bank that are most appropriate for each student, thereby minimizing the number of items needed to ensure an accurate score.

## CONCLUSION

Based on the data analysis, it is concluded that: (1) PhysTCriTS is developed in the form of reasoning multiple choices on two sets, each with 30 items and 8 anchor items. The critical thinking skills test includes the elementary, classification, basic support, fluency, advance clarification, and strategy and tactics of sub-materials of Elasticity, Hooke's Law, Static Fluids, Temperature and Heat, and Optical Devices; (2) PhysTCriTS is eligible and qualified to be used as an instrument to measure Physics critical thinking skills of $10^{th}$ grade students according to the following findings: i) content validation as evidenced by the expert judgment and Aiken's V score of critical thinking skills tests is 0.67 -1.00; ii) empirical validation shown from the result that all of the test items of critical thinking skills fit the PCM, that is, in the range of 0.86 to 1.20; iii) he difficulty indexes obtained spread from -0.75 to 1.30, respectively, or lie between -2.0 and 2.0; therefore, they are categorized as good items; and iv) the entire PhysTCriTS items are reliable based on the values of the information function and SEM; and (3) PhysTCriTS is eligible and qualified to be installed into CAT.

## ACKNOWLEDGMENTS

The authors would like to thank DP2M (Directorate of Research and Community Service) for funding this research.

## REFERENCES

Alfallay, I. S. (2018). Test specifications and blueprints: Reality and expectations. *International Journal of Instruction, 11*(1),195-210. doi: 10.12973/iji.2018.11114a

Alismail, H. A., & Mcguire, P. (2015). 21st century standards and curriculum. *Current Research and Practice, 6*(6), 150-155.

Arif, M. (2014). Penerapan aplikasi anates bentuk soal pilihan ganda. *Jurnal Ilmiah Educatic, 1*(1), 2407-4489.

Azwar S. (2017) *Penyusunan skala psikologi, edisi kedua*. Yogyakarta: Pustaka Pelajar.

Bond, T. G., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences.* New York: Routledge.

Brookhart, S. (2010). *How to assess higher order thinking skills in your classroom*. United States of Amerika: ASCD Member Book.

Cottrell, S. (2011). *Critical thinking skills: Developing effective analysis and argument.* Palgrave Macmillan.

Ennis, R. H., & Weir, E. (1985). *The ennis-weir critical thinking essay test. Test manual, criteria, scoring sheet an instrument for teaching and testing*. United States of America: Midwest Publications.

Finkelman, M. D., Kim, W., Weissman, A., & Cook, R. J. (2014). Cognitive diagnostic models and computerized adaptive testing: Two new item-selection methods that incorporate response times. *Journal of Computerized Adaptive Testing*, *2*(4), 59-76. https://doi.org/10.7333/1412-0204059.

Fischer, A. (2009). *Beprikir kritis*. Jakarta: PT Erlangga.

Hadi, S., & Handhika, J. (2015). Pembelajaran fisika menggunakan modul berbasis scientific approach bermuatan pendidikan karakter pada materi termodinamika. *Prosiding Seminar Nasional Fisika Dan Pendidikan Fisika*, *6*(1), 101-115.

Hadi, S, & Haryanto. (2012). *Hasil belajar berbantuan komputer (Computerized Adaptive).* Paper presented in Seminar Membangun Strategi Evaluasi yang Kredibel untuk Ujian Sekolah dan Ujian Nasional, Universitas Negeri Yogyakarta.

Hambleton R. K., & Swaminathan, H. (1991). *Fundamentals of item response theory*. California: SAGE.

Hosseini, M., Morteza, S., & Toroujeni, H. (2017). Replacing paper-based testing with an alternative for the assessment of Iranian undergraduate students: Administration mode effect on testing performance. *International Journal of Language and Linguistics*, *5*(3), 78-87. https://doi.org/10.11648/j.ijll.20170503.13.

Huang, C. Y., Tung, L. C., Chou, Y. T., Chou, W., Chen, K. L., & Hsieh, C. L. (2018). Improving the utility of the fine motor skills subscale of the comprehensive developmental inventory for infants and toddlers: A computerized adaptive test. *Disability and rehabilitation*, *40*(23), 2803-2809.

Husamah, Fatmawati, D., & Setyawan, D. (2018). OIDDE learning model: Improving higher order thinking skills of biology teacher candidates. *International Journal of Instruction*, *11*(2), 249-264. https://doi.org/10.12973/iji.2018.11217a.

Ibrahim. (2007). *Pengembangan kemampuan berpikir kritis dan kreatif siswa smp dalam matematika melalui pendekatan advokasi dengan penyajian masalah open-ended*. Tesis. Tidak dipublikasikan. Bandung: Universitas Pendidikan Indonesia.

Istiyono, E., Mardapi, D., & Suparno. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (PhysTHOTS) peserta didik SMA. *Jurnal Penelitian dan Evaluasi Pendidikan*, *14*(1), 1-12.

Istiyono, E., Mardapi, D., & Suparno. (2014). Effectiveness of reasoned objective choice test to measure higher order thinking skills in physics implementing of Curriculum 2013. *Proceeding of International Conference on Educational Research and Evaluation (ICERE) in Graduate School Yogyakarta State University*, 79-87

Istiyono, E. (2017). The analysis of senior high school students' physics HOTS in Bantul District measured using PhysReMChoTHOTS. *AIP Conference Proceedings 1868*, 070008. https://doi.org/10.1063/1.4995184.

Linacre, J. M. (2006). *Winstep: Rasch-model computer programs.* Chicago: Winsteps.

Linden, W. J., & Humbleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Lima, E., Teixeira-Salmela, L. F., Magalhães, L. C., Laurentino, G. E., Simões, L. C., Moretti, E., ... & Lemos, A. (2018). Measurement properties of the Brazilian version of the motor assessment scale, based on rasch analysis. *Disability and rehabilitation*, 1-6.

Mardapi, D. (2017). *Pengukuran, penilaian, dan evaluasi pendidikan (edisi kedua)*. Yogyakarta: Parama Publishing.

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for ratingscale data [Computer software]*. Chicago: Scientific Software.

Retnawati, H. (2014). *Teori respon butir dan penerapannya: untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Parama Publishing.

Verhelst N. D., & Verstralen. H. H. F. M. (2008). Some considerations on the partial credit model. *Psicologica*, *29*(2), 229-254.

Widhiarso, W. (2010). Prosedur uji linieritas pada hubungan antar variabel. Retrieved from http://wahyupsy.blog.ugm.ac.id/2010/08/03/prosedur-uji-linieritas-pada-hubungan-antar-variabel/. Diunduh tanggal 15 Maret 2017.

Whiley, D. W., Witt, B., Colvin, R. M., Arrue, R. S., & Kotir, J. (2017). Enhancing critical thinking skills in first year environmental management students: a tale of curriculum design, application and reflection. *Journal of Geography in Higher Education*, 1-20. https://doi.org/10.1080/03098265.2017.1290590.

Živkovic, S. (2016). A Model of critical thinking as an important attribute for success in the 21st century. *Procedia-Social and Behavioral Sciences*. *232*(1), 102-108. https://doi.org/10.1016/j.sbspro.2016.10.034.