



# European Journal of Educational Research

Volume 9, Issue 1, 91 - 101.

ISSN: 2165-8714

<http://www.eu-jer.com/>

## Developing of Computerized Adaptive Testing to Measure Physics Higher Order Thinking Skills of Senior High School Students and its Feasibility of Use

**Edi Istiyono \***

Universitas Negeri Yogyakarta,  
INDONESIA

**Wipsar Sunu Brams**

**Dwandaru**  
Universitas Negeri Yogyakarta,  
INDONESIA

**Risky Setiawan**

Universitas Negeri Yogyakarta,  
INDONESIA

**Intan Megawati**

Universitas Negeri Yogyakarta,  
INDONESIA

*Received: September 9, 2019 • Revised: October 29, 2019 • Accepted: November 27, 2019*

**Abstract:** The Computer has occupied a comprehensive coverage, especially in education scopes, including in learning-teaching processes, testing, and evaluating. This research aimed to develop computerized adaptive testing (CAT) to measure physics higher-order thinking skills (HOTS), namely PhysTHOTS-CAT. The Research Development used the 4-D developmental model carrying the four phases of define, design, development, and dissemination (4D) developed by Thiagarajan. This testing instrument can give the item test based on the student's abilities. The research phases include (1) needs analysis and definition, (2) development design (3) development of CAT and assemble the test items into CAT, (4) validation by experts, and (5) feasibility try-out. The findings show that PhysTHOTS-CAT is valid to measure physics HOTS of the 10th-grade students of Senior High School according to 82.28% of teachers and students assessment on PhysTHOTS-CAT content and media. Therefore, it can conclude that PhysTHOTS-CAT can be used and feasible to measure physics HOTS of the 10th-grade students of the Senior High School.

**Keywords:** *Computerized adaptive testing, HOTS, partial credit model, item response theory.*

**To cite this article:** Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). 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. <https://doi.org/10.12973/eu-jer.9.1.91>

### Introduction

Physics is a subject felt as difficult and frightful for most students. The primary reason is a large number of mathematical formulas involved in physics learning. Besides, the methods adopted by the physics teachers as if implanted that physics concepts are merely collections of formulas, therefore, students persist that they must memorize all those formulas. This condition leads to the development of lower-order thinking skills (LOTS) rather than higher-order thinking skill (HOTS). Mathematics and physics are subjects that are a burden for students because students often have difficulty mastering it (Johnson & May, 2008). Students experiencing difficulties in learning physics need to get attention because these problems can prevent students from getting good results in Physics Learning (Koponen, Mantyla, & Lavonen, 2002). Conceptually, the teachers also feel difficulties in giving students an understanding of physics (Ekici, 2016; Angell, Guttersrud, Henriksen, & Isnes, 2004). A teacher, to be successful, must be trained in each of the necessary dimensions -knowledge, abilities, and relationships (Demkanin, 2018). PISA (Program for International Student Assessment) ranking in the field of Science, Indonesia is ranked 62 out of 70 countries (OECD, 2018). Indicates a low level of potential and ability of students in science so that a breakthrough is needed to improve the ability of physics.

HOTS on physics means the thinking levels of the cognitive domain, well known as Bloom's taxonomy, that covers analyzing, evaluating, and creating (Anderson & Krathwohl, 2010). Macro scoring tends to use samples in analyzing the program and its impacts; the program is called curriculum (Setiawan, 2019). High-order thinking abilities can be developed through the instruction also its assessment. Assessment is conducted to find out how far the students successfully receive the knowledge given by the teacher. A good test requires to construct and proper assessment management. The classic paper-based assessment has many weaknesses in its application. By developing a test-based grading system that can adapt to the ability of students to provide the best solution in the field of measurement. The underdeveloped system combines adaptive assessment based on high order thinking that can measure HOTS capabilities with adaptive computer systems. Evaluation commonly uses regular paper and pencil testing (Cisar, Radosav, Markoski, Pinter, & Cisar, 2010). By the advances of technology, traditional paper, and pencil testing (PPT)

\* **Corresponding author:**

Edi Istiyono, Graduate School, Universitas Negeri Yogyakarta, Colombo Street No.1 Yogyakarta, INDONESIA 55281. ✉ [edi\\_istiyono@uny.ac.id](mailto:edi_istiyono@uny.ac.id)

© 2020 The Author(s). **Open Access** - This article is under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>). 

has experienced a decline due to the length of time in administrating the test and giving feedbacks (Boo & Vispoel, 2012). On the other hand, the use of the computer as the medium for testing has been much developed; for example, the language testing of English (Jamieson, 2009) and Semai (Alwi, Mehat, & Arshad, 2016).

The use of the computer has occupied a comprehensive coverage of scopes, such as in education field. It is convenient to be used in learning-teaching processes, testing, and assessing. (Bennett, 2012; Pommerich, 2004; Fadzil, 2018). Students find a lot of autonomy with the availability of the computer and computer-based testing through computer-based education (CBE), computer-based testing (CBT), computer-based English language testing (CBELT), computer-aided/assisted assessment (CAA), computer-adaptive language testing (CALT), or computerized adaptive testing (CAT). CAT proposes a logical, effective, and efficient measurement in measuring students' abilities (Cella, Gershon, Lai, & Choi, 2007). CAT also optimizes items are managed and can produce the most significant information in a measurement of the ability of test-takers (Haley et al., 2011). CAT generally requires fewer items than long-form instruments and can achieve the same precision (Cella et al., 2007)

CAT is a testing system using a computer that provides the test items that are follow the students' abilities. An adaptive test has five characteristics, they are: (a) it builds upon a bank of test items completed with their statistical characteristics; (b) the test facilitates the test taker to select a starting point from the item bank so not all test takers begin with the same item; (c) test scores can be derived from different test items given to different test takers; (d) the selection of the next items is based on the test taker's responses for the earlier item, if the test taker answer the question correctly the next item will be more difficult, if the answers are wrong the next question will be less complicated. This system was done by employing the Item Response Theory (IRT). (Baker, 1983; Weiss, 2011); (e) a test ends when certain terminating criteria achieved even with the different number of an item done by each individual.

A test package that is inputted into the CAT will be capable of measuring the range of the test-takers' abilities (Lord, 1980). It also more efficient for student testing compared to the conventional ways that require a considerable length of time. Lord also believes that the length of the test can shorten without losing the accuracy of the measurement. The gains of the earlier items determine the sequence of the items displayed on the monitor. It is why the test can shorten without decreasing the accuracy of measurement; therefore, CAT can estimate students' ability levels in a shorter time than other test methods. CAT becomes highly popular for a number of advantages, they are: (a) has high test security (Winarno, 2012); (b) the test is given on request; (c) needs no answer sheets; (d) display items that is related to the test taker's initial ability; (e) the level of test standardization is higher than traditional tests; (f) the items' difficulty slides down to easier items when the test taker's ability level is lower and slides up when the test taker's ability is higher; (g) has a flexibility in selecting the items; (h) supervision time is shorter; and (i) instant reporting and accurate measurement.

In order to measure students HOTS on physics, the test items installed in CAT should be able to measure students HOTS. The multiple-choice test modified with rationals regarded as a further development of the test type, intended to measure students' abilities in all the cognitive levels, particularly the higher-order thinking levels. A good test is one that can accurately measure the test takers' skills, in which the test difficulty index is fitted to the test takers' abilities. Moreover, a good test should consider the steps of completing the test items. An innovative testing method could determine the successfulness of the measurement of the students' abilities; thus, the development of an instrument to measure HOTS in physics by adopting CAT (PhysTHOTS-CAT) need to conduct.

## Methodology

### *Research Goal*

This research aimed to develop CAT to measure 10<sup>th</sup> students' HOTS on physics. The developed CAT will be called as PhysTHOTS-CAT. As the research methodology, the study used the 4-D developmental model carrying the four phases of define, design, development, and dissemination (4D) developed by Thiagarajan (Thiagarajan, Semmel, & Semmel, 1974). The define phase involved the analyses of needs and the definition of the CAT media that is valid, feasible, and efficient to measure HOTS on physics. The design phase consisted of planning the development of the CAT and its scoring system using PCM, which are in line with the needs. The development phase includes the constructing of the CAT, installing the HOTS test items into CAT, and validating the CAT to the experts on learning media. The dissemination phase encompassed the determination of try-out subjects (Senior High School students) and the conduct of the try-out.

### *Sample and Data Collection*

The sample in this research was 300 10<sup>th</sup> grade students of the Senior High School in Yogyakarta. The item response theory approach, if a sample size of 300 meets the requirements for analysis using a partial credit model. The sample chosen was in class 10<sup>th</sup> grade because the aspects in the problem were relevant to the HOTS level at the level of; analysis, synthesis, evaluation, and creating. The data was obtained through a systematic sampling method which selected test participants from 4 schools, one school consisting of 90 to 100 students in Yogyakarta Province. Data was collected at each school in a computer laboratory room with alternating systems according to class groups. The

systematic stratified random sampling where researchers take students randomly based on high, medium, and low levels in HOTS. A random process carried out after the researcher detects or records students who have high, medium, and low abilities based on the school that is the target of this research. The instrument used in this study was a valid, feasible, useful assessment sheet and a test designed to develop CAT. Measurement experts validated measurement experts and physics learning experts validated the assessment sheet used to assess valid, feasible and effective before being used to assess the developed CAT while the tests developed.

*Analyzing of Data*

Test scores are analyzed using the partial credit model (PCM), which is an extension of the one-parameter model. The PCM presents the categories of the test taker's answers in a polytomous scale, an extension of the dichotomous-the working principles of the developed CAT based on those by Winarno (2012). In the beginning, the test takers gave items with an initial standard level of difficulty, which is close to zero. If they can answer it correctly (at scales 3 and 4) on the polytomous model, they will be given items with higher levels of difficulty; and if they answer wrongly (at scales 1 and 2), they will be given items with lower levels. The CAT software uses an algorithm system to presents items that are in line with the test takers' abilities. The algorithm is used to decide the next item to be selected based on the test taker's response to the previous item. In the test administration, the selection of items uses the IRT theory (Reckase, 2009), logics, and simple statistics.

CAT is to stop when the stopping rule is achieved. There are three criteria in determining the stopping rules in the CAT with HOTS items they are: (a) when the time is used up, (b) when the accuracy level of the estimate of the test taker's abilities is achieved, or (c) the differences among the standard error of measurement (SEMs) in the item repetition is markedly low ( $SEM < 0.01$ ), as shown in Figure 1.

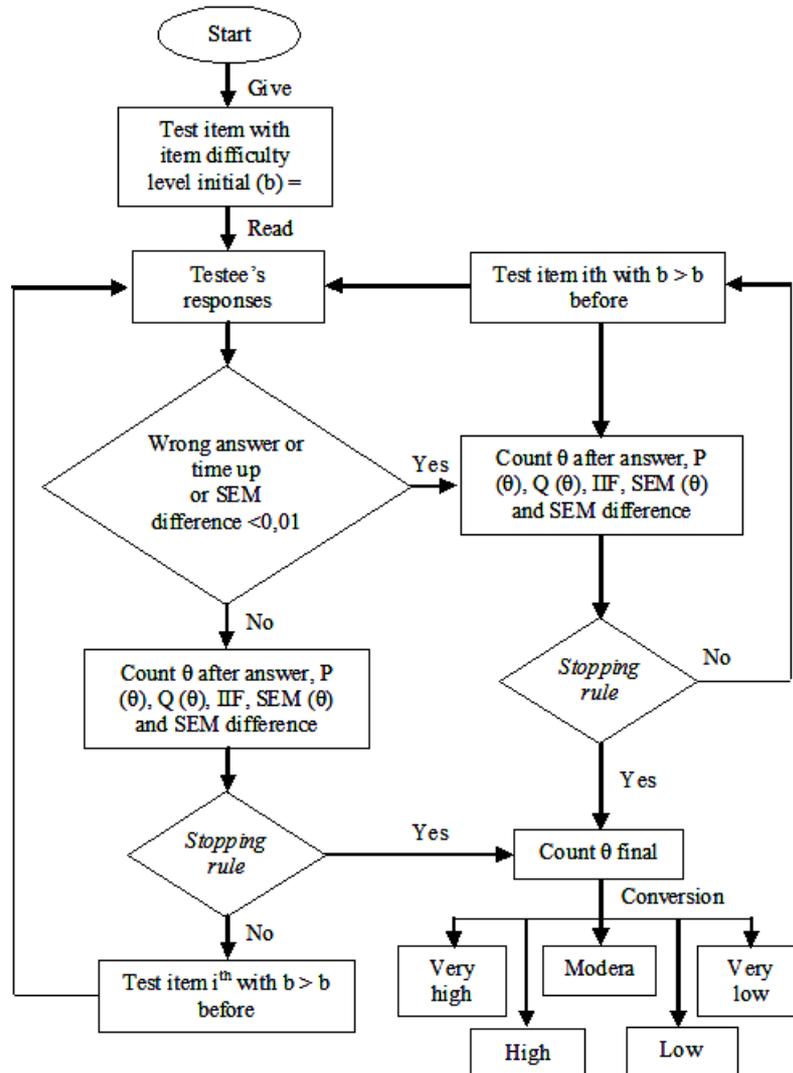


Figure 1: Stopping rule for test items in CAT

The test items that were inputted into the CAT was physics HOTS test items set developed by Megawati and Istiyono (2017). The CAT procedure in figure 1 starts with questions based on the level of difficulty. After the initial assessment begins on one question, then the response will result in the level of difficulty item. After that, the standard measurement process error of measurement (the criteria must be <0.01). The results of the procedure calibrated with incorrect answers and measurement errors. The next process is the calculation of the measurement error variance through the measurement information function. After calibration, the system will calculate capabilities (theta) and group them in five categories of scale, namely, very high, high, moderate, low, and very low. All items were valid test items as shows by their fit against the PCM; had a favorable difficulty index (well between -2.0 and 2.0), and had a reliability measure as evident from the information function and SEM (Azwar, 2015). The physics HOTS test items were suitable for students with the abilities level  $\theta \geq -1.4$ , in moderate and high categories.

Validation on the CAT media was conducted to find out the degree of the feasibility of the media being developed. The media was computerized adaptive testing (CAT) software. Data analyses for the media feasibility adapted the data analysis technique developed by Winarno (2012). The validation sheets from the media experts were converted into five interval scores of very good, good, moderate, poor, and very poor.

For the feasibility measure of the CAT development, questionnaires were given to the teachers and students to find out their responses towards the implement ability of the CAT test. Responses to the questionnaires were analyzed descriptively. Data analyses were done by counting the scores obtained from all the aspects being scored and subsequently calculated by way of the following formula.

$$N = \frac{k}{N_k} \times 100\% \dots\dots\dots (1)$$

Notes:

- N = Percentage of aspect feasibility
- k = Data collection score
- N<sub>k</sub> = Highest total score

In the equation used to measure students' abilities ( $\theta$ ), the probability of giving correct answers of scales 3 and 4 is P<sub>i</sub> ( $\theta$ ), the probability of giving wrong answers of scales 1 and 2 is Q<sub>i</sub> ( $\theta$ ), information function I<sub>i</sub> ( $\theta$ ), and SEM ( $\theta$ ). Use of the logistic model 1 PL applies the Rasch model so that in order to determine ( $\theta$ ), P<sub>i</sub> ( $\theta$ ), Q<sub>i</sub> ( $\theta$ ), I<sub>i</sub> ( $\theta$ ) and SEM ( $\theta$ ) the following equation is used:

- a. The equation to determine students' HOTS abilities ( $\theta$ ),

$$\theta = b_i + \frac{1}{D\alpha_i} \ln\left(0.5\left(1 + \sqrt{1 + 8c_i}\right)\right) \dots\dots\dots (2)$$

Notes:

- $\theta$  : students' HOTS abilities (-3 <  $\theta$  < +3)
- $b_i$  : Item difficulty index <sup>i</sup>th
- D : Scaling factor (count as 1,7)
- $\alpha_i$  : Item differentiating index <sup>i</sup>th (for logistic 1PL, model value of =  $\alpha_i$  1)
- $c_i$  : Pseudo guessing index on item <sup>i</sup>th (for logistic 1PL and 2PL, model value of )  $c_i = 0$

- b. Equation do determine the probability of correct answering P<sub>i</sub>( $\theta$ )

$$P_{ig}(\theta) = \frac{\exp\left[\sum_{j=0}^x \theta - b_j\right]}{\sum_{h=0}^m \exp\left[\sum_{j=0}^x \theta - b_j\right]} \text{ where } j = 1,2,3\dots m+1 \dots\dots\dots (3)$$

- c. Equation do determine the probability of wrong answering Q( $\theta$ )

$$Q(\theta) = 1 - P_{ig}(\theta) \dots\dots\dots (4)$$

- d. Equation do determine information function I<sub>i</sub>( $\theta$ )

$$I_i(\theta) = Q_i(\theta)P_i(\theta) \dots\dots\dots (5)$$

e. Equation do determine information function SEM

$$SEM(\theta) = \frac{1}{\sqrt{\sum_{i=1}^N I_i(\theta)}} \dots\dots\dots (6)$$

Each test taker's higher-order thinking ability coming out of CAT is then categorized under five abilities levels adopted by Istiyono (Istiyono, Dwandaru, Megawati & Ermansah, 2017), that are: very low, low, moderate, high, and very high.

## Findings / Results

### *The Computerized Adaptive Test (CAT)*

The developed CAT has the following main components: opening page, administering page, teacher's page, and student's page. First, the opening page displays the media title, CAT logo and developer logo, Yogyakarta State University (YSU). At the bottom of the page, the login button for the administrator, teacher, and student are provided. When the button is clicked, the login page is appeared. The particular user has to type the username and password in order to log in to the system. If a user is not registered yet, they have to register themselves to the administrator.

Second, the administration page contains Home, Data Master, and Report Recapitulation menus. The Data Master contains the Administrator Setting, Teacher, Student, and Item Bank. The Administrator Setting is a menu to register any administrators or attendants and obtain the username and password. In the Administrator page, there is also the Report Recapitulation, the last menu on the Administrator page that provides the test results. Before entering the Report Recapitulation page, the user must first select the type and date of the test. Test results present the history of test results of each test taker. Similarly with the administrator setting, the teacher and student menus are the operation pages for teachers and students. The item bank contains 62 valid and reliable HOTS test items, that is previously developed by Megawati and Istiyono (Istiyono et al., 2017) and already proven to be valid and reliable. Valid and reliable become the main criteria because an instrument can measure the ability in accordance with the objectives and show the consistency of an instrument. The administrator is able to add or delete any teachers, students or items and set the item mode by determining when the test begins and ends.

Third, the Teacher menu contains Home, Item Bank, and Test Result menus. Each menu has the same function, except for the Setting menu. The setting menu is only for the administrator to make any addition or deletion. The last page is the Student page. Only test takers who are registered in the administrator's database can be logged in. The page contains home, test list, and Test Result, and the test List menu is to start the test. The test display consists of a test item, time, and answer button. After the test is over, the test takers can view the results in "All" or "Custom" menu. The "All" menu presents results in details. The test takers who want to know the simple version of the test result can select "Custom."

### *CAT Validation*

The validation of the CAT is done to know the feasibility of the developed media. Validation of the PhysTHOTS-CAT is conducted by involving four media experts and five teachers. The validation results of the PhysTHOTS-CAT were 38.5 for the display which is within the very good category, 16.5 for the media feasibility which is within the good category, and 55.0 for the total which is within the very good category. According to the experts, the developed PhysTHOTS-CAT is rated very well and feasible, but numbers of suggestions are obtained from the validation process. The suggestions are: (a) on the front page, information about the study program that develops the test need to be provided (see Figures 2); (b) the "Answer" button, which is initially on the right side, need to be moved to the left side for the sake of efficiency (see Figures 3); and (c) the test-result display should be easier to understand by adding columns for true and false answer also the reasons (see Figures 4).

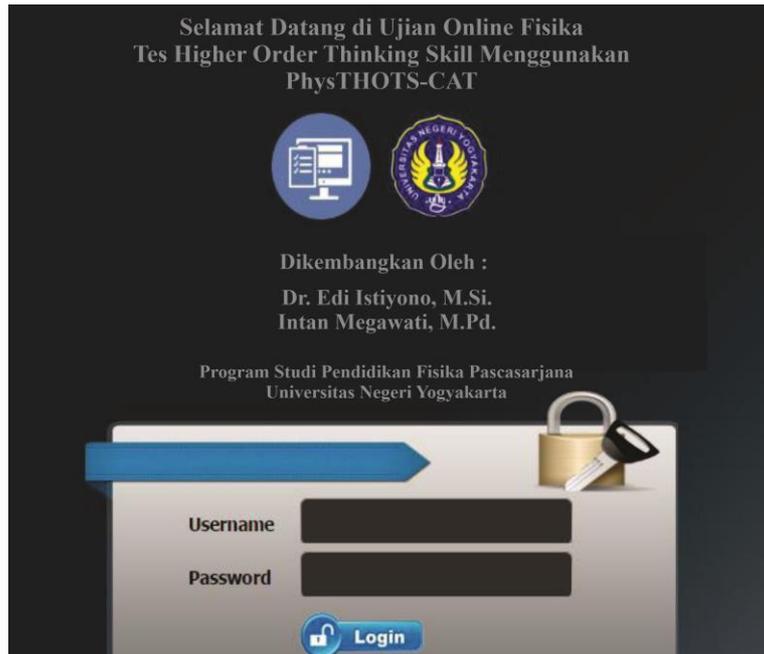


Figure 2: Front page of programe display

1
6.1.6
59:52

Two densely-populated substances have the same period , aluminium and copper ( $C_{al} = \frac{900J}{Kg}^{\circ}C$  dan  $C_{tem} = \frac{390J}{Kg}^{\circ}C$  ) , on a trial of temperature changes and the heat transfer to the two solid substances heated for 15 minutes to occur temperature changes around 80°C. If asked to formulate a hypothesis, which one fits your hypothesis?

- A.  Aluminum absorbs the same heat with copper when heated at the same time
- B.  Aluminum absorbs the same heat with copper when heated at the same temperature
- C.  Aluminum absorbs the heat equal to copper when heated because it has the same mass
- D.  Aluminium absorbs color Greater than on copper when heated at the same time
- E.  Aluminium absorbs color less than on copper when heated at the same time

The reason:

- A.  To raise the temperature of 1°C The temperature of substances affected by the temperature change substances
- B.  To raise the temperature of 1°C substances temperature affected by mass substances
- C.  To raise the temperature of 1°C The temperature of substances affected by the capacity of substances
- D.  To raise the temperature of 1°C The temperature of substances affected by the heat type of substances
- E.  To raise the temperature of 1°C the temperature of substances affected by Heating time

ANSWER

Figure 3: The "Answer" button display



Figure 4: The test-result display

### CAT Try-out

From the limited-scope try-out, no major obstacle has been found in the test takers when they are doing the test. In the try-out, the CAT has been found to function well. One disturbing situation is when test takers do not read the directions carefully so that they push the backspace key and they cannot repeat the test. Another difficulty occurs when the Internet network is not maximal so that test takers have to pause and fail to continue to the next items. Some respondents reported that they are not so used to the multiple-choice test version with rationals and find it difficult to handle this multiple-choice HOTS version. One main obstacle is the situation where the Internet network is slow. Notwithstanding, use of the CAT is categorized as feasible so that it can be continued to the phase of the field try-out.

In the field try-out, the research instruments are CAT software and questionnaires for the teachers and students. All the research instruments have been checked for validity and reliability and are feasible for collecting data. The field try-out is aimed at measuring the students' HOTS using the CAT test items and finding the feasibility measure of the administration of the developed CAT.

Students' HOTS abilities are found to be of the moderate category with 0.04 as the lowest ability and 0.61 as the highest. Within the 60 minutes of the test time, the maximum number of items completed is 23. Results of HOTS testing using the PhysTHOTS-CAT are presented in Table 1.

Table 1. Results of the PhysTHOTS-CAT

No	ID	Class	Teacher	Ability	Item Number	Score	Time
1	16954	XI MIA 3	I M	0.04	14 items	50.67	58:21
2	16957	XI MIA 3	I M	0.15	19 items	52.50	54:31
3	16974	XI MIA 3	I M	0.38	22 items	56.33	56:03
4	16987	XI MIA 3	I M	0.10	14 items	51.67	38:51
5	17000	XI MIA 3	I M	0.16	19 items	52.67	53:15
6	17001	XI MIA 3	I M	0.16	20 items	52.67	55:28
7	17016	XI MIA 3	I M	0.29	22 items	54.83	59:37
8	17017	XI MIA 3	I M	0.11	15 items	51.83	43:35
9	17043	XI MIA 3	I M	0.33	21 items	55.50	59:51
10	17061	XI MIA 3	I M	0.20	20 items	53.33	59:35

One example of the score details of the PhysTHOTS-CAT and the score history is presented in Figures 5 and 6.

305/2017 Rekap Laporan Test

Nama Peserta : Siti F  
 Kelas : XI MI  
 Nama Test : HOTS FISIKA  
 Waktu Test : Selasa, 21 Feb 2017  
 Pengerjaan Test : 59 Menit : 45 Detik  
 Nilai : 60.17



Item	KD	b	bi(1)	bi(2)	bi(3)	Poin	θ awal	θ akhir	P <sub>ni</sub> (θ)				P <sub>i</sub> θ	Q <sub>i</sub> θ	IIF	SE(θ)	Selisih SE
									1	2	3	4					
1	4.2.6	0.02	-1.44	3	1.63	4	0.02	0.02	0.180	0.770	0.040	0.010	0.010	0.990	0.01	11.33	11.33
2	5.1.3	0.04	0.36	-0.34	0.1	4	0.02	0.04	0.270	0.190	0.280	0.260	0.260	0.740	0.19	2.233	9.097
3	4.3.4	0.08	-0.81	1.3	0.75	4	0.04	0.08	0.230	0.540	0.150	0.080	0.080	0.920	0.07	1.928	0.307
4	6.3.2	0.1	0.52	0.95	-1.16	3	0.08	0.1	0.350	0.230	0.090	0.330	0.090	0.910	0.09	1.676	0.25
5	4.3.2	0.11	-1.1	1.71	0.95	4	0.1	0.11	0.190	0.630	0.130	0.050	0.050	0.950	0.05	1.56	0.116
6	5.1.8	0.12	0.94	-0.7	0.11	4	0.11	0.12	0.290	0.130	0.290	0.290	0.290	0.710	0.21	1.275	0.285
7	6.3.3	0.14	1.01	-0.63	0.05	4	0.12	0.14	0.310	0.130	0.270	0.290	0.290	0.710	0.21	1.1	0.175
8	6.2.4	0.15	0.73	-0.1	-0.17	4	0.14	0.15	0.310	0.170	0.220	0.300	0.300	0.700	0.21	0.981	0.119
9	5.1.1	0.16	0.48	-0.29	0.29	4	0.15	0.16	0.260	0.190	0.290	0.260	0.260	0.740	0.19	0.902	0.079
10	5.2.6	0.2	0.07	1.47	-0.93	4	0.16	0.2	0.310	0.330	0.090	0.270	0.270	0.730	0.2	0.837	0.065
11	6.1.6	0.25	1.12	0.56	-0.94	4	0.2	0.25	0.390	0.160	0.110	0.340	0.340	0.680	0.22	0.777	0.06
12	5.2.5	0.29	1.04	0.65	-0.8	3	0.25	0.29	0.380	0.170	0.120	0.330	0.120	0.880	0.1	0.755	0.022
13	5.1.6	0.3	-0.02	1.68	-0.76	4	0.29	0.3	0.270	0.370	0.090	0.260	0.260	0.740	0.19	0.717	0.038
14	6.1.2	0.33	0.51	-0.95	1.41	4	0.3	0.33	0.180	0.150	0.510	0.170	0.170	0.830	0.14	0.693	0.024
15	4.2.7	0.38	-1.48	1.23	0.88	4	0.33	0.38	0.090	0.550	0.230	0.130	0.130	0.870	0.11	0.675	0.018
16	6.1.5	0.39	0.5	0.16	0.51	4	0.38	0.39	0.250	0.220	0.280	0.240	0.240	0.760	0.18	0.649	0.026
17	6.2.1	0.4	1.32	-0.89	0.78	4	0.39	0.4	0.260	0.100	0.380	0.250	0.250	0.750	0.19	0.625	0.024
18	5.1.9	0.42	-1.12	1.18	1.32	3	0.4	0.42	0.120	0.540	0.250	0.100	0.250	0.750	0.19	0.603	0.022
19	6.1.3	0.44	0.32	0.21	0.8	4	0.42	0.44	0.230	0.250	0.310	0.210	0.210	0.790	0.17	0.586	0.017
20	5.2.3	0.45	1.01	-0.37	0.7	4	0.44	0.45	0.260	0.150	0.330	0.260	0.260	0.740	0.19	0.567	0.019
21	6.3.4	0.58	1	0.66	0.08	1	0.45	0.58	0.370	0.210	0.170	0.250	0.370	0.630	0.23	0.547	0.02
22	5.1.10	-0.02	0.38	-0.38	0.57	3	0.58	-0.02	0.120	0.140	0.370	0.370	0.370	0.630	0.23	0.529	0.018
23	4.1.4	0.61	1.84	-0.46	0.46	3	-0.02	0.61	0.650	0.100	0.160	0.100	0.160	0.840	0.13	0.52	0.009

θ Akhir : 0.61  
 Skor Poin : 60.17

Figure 5: A test score result of a test taker

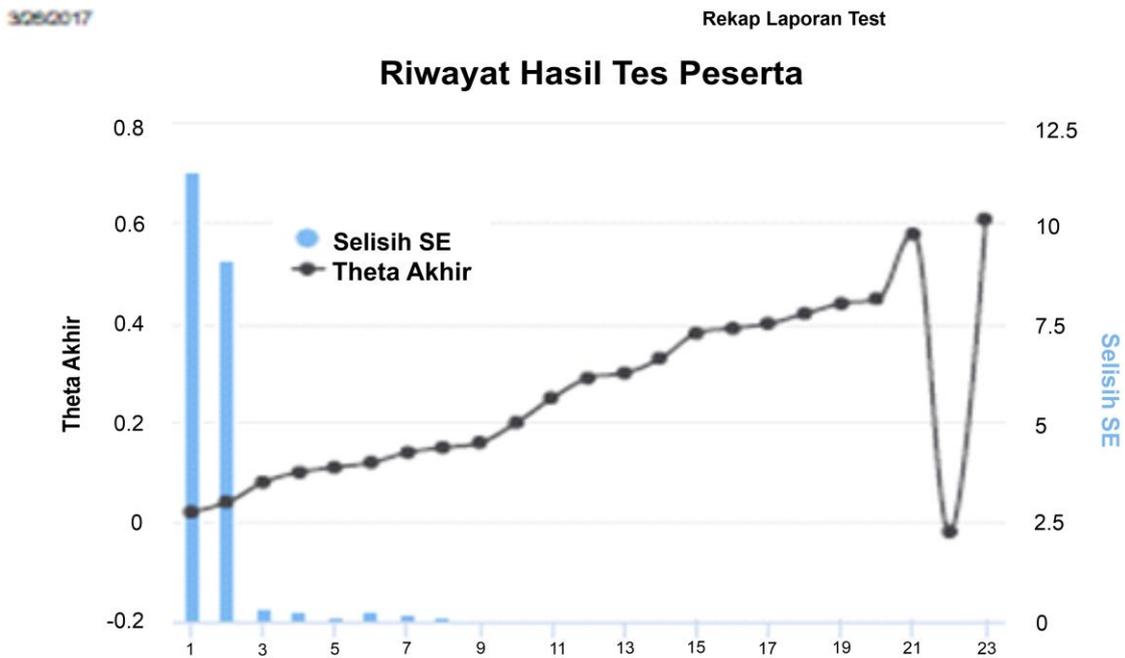


Figure 6: Test score history

In Figures 5 and 6 show that the SEM differences is decreasing. This means that the measurement of student’s HOTS is getting closer to the test taker’s abilities. This particular test taker answers 23 test items in 59 minutes 45 seconds, and achieves a final ability of 0.61, score point of 60.17, and a abilities level within the very high category.

From results of these analyses, it can be concluded that the developed PhysTHOTS-CAT has run as it has been expected. It is considered to be able to measure the test taker’s abilities in individual measures. Furthermore, the percentages of test takers’ abilities obtained from the PhysTHOTS-CAT try-out is shown by Figure 7. Figure 7 shows that the high score category dominated by 33 students and followed in the low category by 30 students. This means that the distribution of students' abilities is diverse and has a spread distribution.

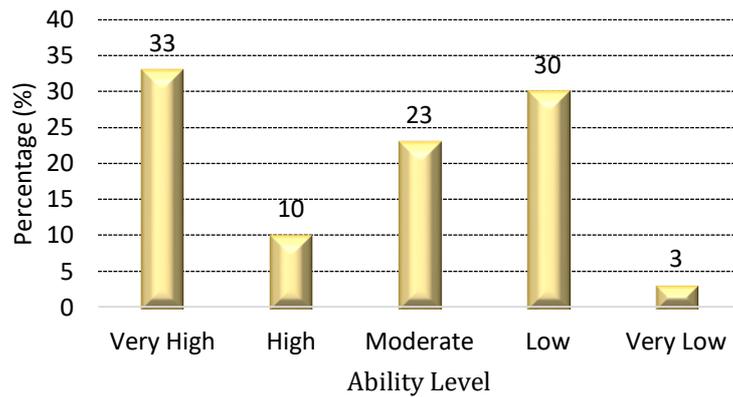


Figure 7: Percentage of Ability levels

#### Feasibility of the use of PhysTHOTS-CAT

The data shows the product feasibility test obtained from 99 students and 10 physics teachers showed a high average score of 82.28. The final phase of the study is to know the feasibility level of the use of *PhysTHOTS-CAT* in measuring students' HOTS. The responses of the teachers and students toward the questionnaires are presented in Table 2.

Table 2. Teachers and Students' Responses of PhysTHOTS-CAT

No	Respondents	Feasibility (%)
1	Students	83.06
2	Teachers	81.50
Average:		82.28

#### Discussion and Conclusion

Technological devices such as a computer, give many opportunities to teachers and students (Feyzioglu, Akpinar & Tatar, 2018). Today, computer technology influences the structure, content, and learning processes are arranged in a curriculum, assessment, measurement, and, evaluation (Dunkel, 1999). It helps students to learn scientific concepts meaningfully. It also can provide media for teachers to assess students' ability. The purpose of this research is to develop a CAT which able to measure physics HOTS, named *PhysTHOTS-CAT*. The physics HOTS is included analyze, evaluate, and create ability. The test had to be able to give an item based on the test taker's ability. Based on expert judgment, *PhysTHOTS-CAT* is rated very good and feasible to be used to measure physics HOTS. *PhysTHOTS-CAT* was able to measure test taker's ability to solve the test well with randomly given test items based on IRT. The test which based adaptive system in giving items could give the more accurate result of the test (Winarno, 2012, Feyzioglu, et al., 2018). The accurate result of the test is very important in getting information or data about student ability (Andrian, Kartowagiran, & Hadi, 2018). A accurate test or valid can attend true information about the skill or ability of student (Wynd, Schmidt, & Schaefer, 2003). In generally, *PhysTHOTS-CAT* could choose and giving an item test to test takers based on their ability and could measure their ability accurately.

Based on Table 1, the test taker's minimum ability was 0.04 with score 50.67, and the maximum ability was 0.61 with score 60.17. From Figure7, it can be seen that the students' HOTS are in the categories of low, moderate, high, and very high. This can be explained by the characteristics of the tests. The results show the distribution of data in the order of categories; very high (33), low (30), moderate (23), high (10), and very low (3). For completing the test items, test takers must be able to find the accurate information about the pictures and present it in the form of mathematical or verbal solutions (Istiyono et al., 2017; Fadzil, 2018). This shown that if test takers can solve many items correctly, it indicated their ability at higher level. High ability is an important variable in physic learning because high ability can determine how far success in next physics learning (Cavallo, Potter, & Rozman, 2004). The difference in the level of ability in physics learning is one of the determining factors for the success or failure of learning physics (Folashade & Akinbobola, 2009).

In Table 2, it can be seen that *PhysTHOTS-CAT* records a feasibility estimate of 82.30 % under the category of "feasible". It can be seen that *PhysTHOTS-CAT* is able to measure the students' abilities in completing a test in accordance with the setting of the test administration which is random, using simple logic, and by IRT polytomous items. Use of this algoarhythm adopts the results of a study conducted by Winarno (2012). Use of a web-based CAT gives advantages in managing data on a vast scale. Analyses of large-scale data give more favorable results (Ahmad, Ishak, Alias & Mohamad, 2015) and, with CAT, such analyses can be carried out more easily. Feasibility of *PhysTHOTS-CAT* shows

that the CAT, which is based on adaptive systems and giving test items that are in line with students' abilities, the developed software gives test results with high accuracy. The CAT development is one of the innovations in testing and assessment, which is in line with the item response theory. Relevant to the research of Ramadan (2019) HOTS instruments developed have characteristics as useful instruments and meet the requirements used to measure. Based on the analyses and results of the study, it can be concluded that the developed PhysTHOTS-CAT can be used as a test instrument consisting of valid and reliable items, validated by experts, and is feasible for measuring physics HOTS of students of 10<sup>th</sup> grade of the Senior High School.

### Suggestions

Referring to the results of the research that each middle school unit is expected to use the CAT program to assess students' HOTS, so that students can change their ability to be better in Higher Order Thinking in Physics. The use of time variables and question bank sorting based on the level of difficulty become a method of measuring students' HOTS abilities.

### Acknowledgements

Great appreciation and gratitude are due to the DP2M (Ministry of Higher Education Research and Community Service) that has provided funding for this research project.

### References

- Ahmad, N. B., Ishak, M. K., Alias, U. F., & Mohamad, N. (2015). An approach for e-learning data analytics using SOM clustering. *International Journal of Advances Soft Computing and its Application*, 7(3), 94-112.
- Angell, C., Guttersrud, O., Henriksen, E. K. & Isnes, A. (2004). Physics: Frightful, but fun, pupils' and teachers' views of physics and physics teaching (Electronic version). *Science Education*, 88, 683-706.
- Alwi, A., Mehat, M., & Arshad, N. I. (2016). E-semat teaching portal (ESTP): A preliminary study in assisting the teaching of Bahasa Semat. *International Journal of Advances Soft Computing and its Application*, 8(1), 57-69.
- Anderson, L. W. & Krathwohl, D.R (Eds). (2010). *Kerangka landasan untuk pembelajaran, pengajaran, dan asesmen: revisi taksonomi pendidikan Bloomian* [A framework for learning, teaching and assessment: a revision of the Bloomian educational taxonomy]. Yogyakarta, Indonesia: Pustaka Pelajar.
- Andrian, D., Kartowagiran, B., & Hadi, S. (2018). The instrument development to evaluate local curriculum in indonesia. *International Journal of Instruction*, 11(4), 922-934. <https://doi.org/10.12973/iji.2016.9115a>
- Azwar, S. (2015). *Reliabilitas dan validitas* [Reliability and validity]. Yogyakarta, Indonesia: Pustaka Pelajar.
- Baker, F. B. (1983). *The basic of Item response theory*. Portsmouth, NH: Heinemann.
- Bennett, R. E. (2012). Inexorable and inevitable: The continuing story of technology and assessment. *The Journal of Technology, Learning and Assessment*, 1(1), 1-23.
- Boo, J. & Vispoel, W. (2012). Computer versus Paper-and-Pencil assessment of educational development: A comparison of psychometric features and examinee preferences. *Psychological Reports* 111, 443-460.
- Cisar, S.M., Radosav, D., Markoski, B., Pinter, R., & Cisar, P. (2010). Computerized adaptive testing of student knowledge. *Acta Polytechnical Hungaria* ,7(4), 139-152.
- Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, 16(SUPPL. 1), 133-141. <https://doi.org/10.1007/s11136-007-9204-6>
- Cavallo, A. M. L., Potter, W. H., & Rozman, M. (2004). Gender differences in learning constructs, shifts in learning constructs, and their relationship to course achievement in a structured inquiry, yearlong college physics course for life science majors. *School Science and Mathematics*, 104(6), 288-300. <https://doi.org/10.1111/j.1949-8594.2004.tb18000.x>
- Depdikbud. (2003). Republic of Indonesia Law Number 20, 2003, about the National Education System. Jakarta, Indonesia: Ministry of Education and Culture.
- Demkanin, P. (2018). Concept formation : Physics teacher and his know-how and know-why. *Journal of Baltic Science Education*, 17(1), 4-7.
- Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology*, 2(2), 77-93.

- Ekici, E. (2016). "Why do i slog through the physics?" understanding high school students' difficulties in learning Physics. *Journal of Education and Practice*, 7(7), 95–107. <https://doi.org/10.1016/j.pec.2014.12.003>
- Fadzil, H. M. (2018). Designing infographics for the educational technology course: Perspectives of pre-service science teacher. *Journal of Baltic Science Education*, 7(1), 8-18.
- Feyzioglu, E. Y., Akpınar, E., & Tatar, N. (2018). Effect of technology-enchanged metacognitive learning platform on accuracy and understanding of electricity. *Journal of Baltic Science Education*, 7(1), 43-64.
- Folashade, A., & Akinbobola, A. O. (2009). Constructivist problem based learning technique and the academic achievement of physics students with low ability level in nigerian secondary schools. *Eurasian Journal of Physics and Chemistry Education*, 1(1), 45–51. <https://doi.org/10.1111/j.1469-8749.2011.04107.x>
- Haley, S. M., Coster, W. J., Dumas, H. M., Fragala-Pinkham, M. A., Kramer, J., Ni, P., ... Ludlow, L. H. (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>
- Istiyono, E., Dwandaru, W. S. B., Megawati, I., & Ermansah. (2017). Application of Bloomian and Marzanoian higher order thinking skills in the physics learning assessment: An Inevitability. In *Proceedings of the International Conference on Learning Innovation (ICLI 2017)* (pp. 136-142). Paris, France: Atlantis Press. <https://doi.org/10.2991/icli-17.2018.26>
- Jamieson, J.M. (2009). Trends in computer-based second language assessment. *Journal of Applied Linguistics*, 25, 228-242.
- Johnson, D., & May, I. M. (2008). *The teaching of structural analysis: A report to the Ove Arup Foundation*. London: UK: The Ove Arup Foundation
- Koponen, I. T., Mantyla, T., & Lavonen, J. (2002). Challenges of web-based education in physics teachers' training. In *Proceedings of ICTE - 2002 (International Conference on Information Communication Technology in Education)* (pp.291-295).
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- OECD (2018). PISA 2015: PISA Results in focus. Retrieved from <https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf>
- Pommerich, M. (2004). Developing computerized versions of paper-and pencil tests: Mode effects for passage-based tests. *The Journal of Technology, Learning and Assessment*, 2(6), 3-44.
- Ramadhan, S., Mardapi, D., Prasetyo, Z.K., & Utomo, H.B. (2019). The development of an instrument to measure the higher order thinking skill in physics. *European Journal of Educational Research*, 8(3), 743-751. doi:10.12973/eu-jer.8.3.743
- Reckase, M.D. (2009). *Multidimensional item response theory (statistic for social and behavioral sciences)*. London, UK: Springer Science and Bussines Media.
- Setiawan, R. (2019). A comparison of score equating conducted using haebara and stocking lord method for polytomous. *European Journal of Educational Research*, 8(4), 1071-1079. doi:10.12973/eu-jer.8.4.1071
- Thiagarajan, S., Semmel, D. D., & Semmel, M. I. (1974). *Instructional development for training teacher of expectional children*. Minneapolis, MN: University of Minnesota.
- Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, 2(1), 1-27.
- Winarno. (2012). Pengembangan computerized adaptive testing menggunakan metode pohon segitiga keputusan [Development of computerized adaptive test using the decision triangle tree method]. *Journal of Education and Educational Evaluation/Jurnal Pendidikan dan Evaluasi Pendidikan*, 16(2), 574-592.
- 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>