

The development of CAT-MARZANO as an assessment media in the industrial revolution 4.0

Y A B Patria¹ and E Istiyono¹

¹Physics Education, Universitas Negeri Yogyakarta, Sleman, Indonesia

Corresponding author: yohan.aurino.2016@student.uny.ac.id

Abstract. In the Industrial Revolution 4.0 (IR 4.0) computers were used as a media of assessment. This study aims to develop a media assessment in the form of CAT-MARZANO as a form of response to the challenges of IR 4.0 in the world of education. The research method used is the R&D method using the 4D model to develop the CAT and Antonio Oriundo models to develop HOTS of Marzano's version test. This research was conducted on 11th grade students in Bantul District. The results of this study found that 136 questions developed were valid and reliable based on the Aiken V index (0.74.1.00) which was assessed by 9 experts. Test fit with PCM and effective for use with a percentage of 82.80%.

Keywords: *assessment, CAT, marzano, test*

1. Introduction

Assessment is an important part in the learning process. Assessment is not only important for educators, but also important for students, schools, parents, and even society. Assessment is one way to find out the learning outcomes of students [1]. A good assessment can be done by collecting accurate evidence related to the achievement of students' learning outcomes and making the class assessment process and the results useful for students, which is able to increase motivation and learning achievement [2].

Test and non-test assessments are the form of assessment. Examples of test assessments are selective tests, initial tests, final tests, diagnostic tests, formative tests, and summative tests. While examples of non-test assessments are observation, report forms, audio-visual techniques, sociometry techniques, and others. Based on the form of the questions, tests can be grouped into two forms, namely objective tests and non-objective tests. Objective tests include multiple choice, short answers, matched, and objective descriptions, while the form of non-objective questions is free description, performance or observation, and portfolio or project [3].

One type of objective test is a two-level objective test that consisting of a choice of answers and a choice of reasons used as a basis for answering [4]. Two-level tests are effective in determining students' conceptual understanding and can also be used as an alternative to assessing and evaluating student achievement [5]. Examples of two-level tests are reasoning multiple choice tests as a form of development of ordinary multiple choice tests that can be used to measure students' higher-order thinking skills [6].

In Indonesia, large-scale assessments are carried out in the form of National Examinations (UN) and carried out with test assessments. The purpose of large-scale assessment is to improve the educational process by monitoring students' learning achievement [7]. The assessment is done using computer media as a form of adjustment to technological developments, as a challenge to the industrial revolution 4.0 in



the world of education. This computer-assisted test is known as a computer-based test or CBT. Random choice of large question banks, innovative and sophisticated item formats, direct feedback, automatic assessment and reporting, question security are some of the advantages of CBT compared to pencil and paper tests (PPT) [8].

Although CBT is better than PPT, CBT still has weaknesses because it only presents questions randomly without taking into account the difficulty of the questions. From this weakness, the Computerized Based Test (CBT) was further developed into the Computerized Adaptive Test (CAT). CAT can present questions randomly while calculating the difficulty level of the questions for the users of the test. CAT is a form of test that is in accordance with students' personalization features [9]. CAT is a test method to improve measurement precision and reduce the number of items [10]. CAT is suitable to measure the students' higher order thinking skills in HOTS of Bloomian [11]. This research developing HOTS in CAT for HOTS of Marzano version. The study aims of this research to develop a media assessment in the form of CAT-MARZANO as a form of response to the challenges of IR 4.0 in the world of education. CAT-MARZANO can be used in line with the IR 4.0 where technology is used to make it easy to do an assessment without losing the essence of that assessment.

2. Research method

This paper uses research and development methods with 4D model to develop computerized adaptive test and Antonio Oriundo model to develop test HOTS Physics of Marzano's version. The 4D model includes the Define, Design, Development, and Disseminate stages. Antonio Oriundo's model includes the Design, Trials, and Assembling the test stages [12]. This research was conducted in Bantul District on grade 11st students as many as 322 students as research subjects. Develop assessment for higher order thinking skills is not easy as lower order thinking skills [13].

2.1. Computerized adaptive test

Computerized Adaptive Test is a type of computer-based test that adapts to the ability level of the participants [14]. CAT provides a promising solution to providing tests with adequate accuracy and general measures to measure differences in individual abilities [15]. In contrast to classical tests, individual assessment becomes a reality through adaptive assessment which gives an opportunity to find out students according to their ability level [16]. This shows that although the questions that are done are different, the things that will be measured are the same.

Computerized adaptive test is a sophisticated method of delivering tests. Nearly 40 years of research has been carried out on this assessment method [17]. CAT can be used to promote more effective learning by testing various skills, knowledge and understanding [18]. Assessment will be more meaningful because the benefits can be felt in the long term both by students and educators.

2.1.1. Define. This define stage is important before starting to develop a product. At this stage the definition of product specifications will be made. In addition, this stage also aims to determine the target users of the assessment media. The developed CAT can be used not only by admin and students, but it can also be used by teachers or other users who need innovation in assessment according to the industrial revolution 4.0.

2.1.2. Design. The design stage starts with creating a story board. Story boards are used as a reference in making website-based assessment media. After the story board is created, the next step is to create a website design.

2.1.3. Development. This development stage is carried out in accordance with advice given by media experts. After the CAT has been made, the questions are ready to be entered into the CAT which can then be used as a media assessment.

2.1.4. Disseminate. This stage is the last stage after the CAT-MARZANO has been completed and is feasible and valid to measure physics higher order thinking skills of Marzano version. Dissemination is done by giving products to teachers in several high schools.

2.2. Higher order thinking skills of Marzano

Marzano identifies 13 higher order thinking skills, namely abstracting, decision making, inductive reasoning, deductive reasoning, comparing, classifying, analysing errors, constructing supporting, analysing perspectives, investigation, problem solving, experimental inquiry, and invention [19].

2.2.1. Design. The items developed are in the form of reasoning multiple choice questions based on indicators in the Marzano taxonomy. These aspects, sub-aspects, and indicators become a reference in the development of HOTS questions. The Marzano indicator is shown in table 1.

Table 1. Aspect, sub aspect, and indicator of marzano.

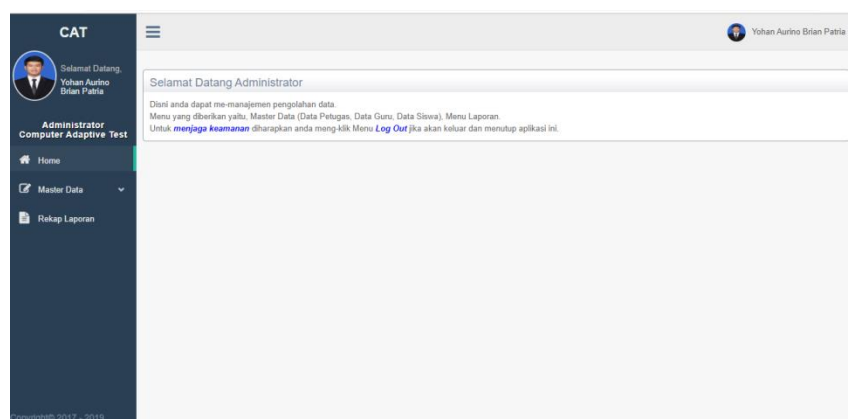
Aspect	Sub Aspect	Indicator of Marzano
Knowledge Recovery	Abstracting	Identifying and articulating the underlying theme or general pattern of information.
	Decision Making	Generating and applying criteria to select from among seemingly equal alternative
Understanding	Inductive Reasoning	Inferring unknown generalization or principles from information or observation
	Deductive Reasoning	Using generalization and principles to infer unstated conclusion about specific information or situations
Analysis	Comparing	Identifying and articulating similarities and differences among items
	Classifying	Grouping things into definable categories on the basis of their attributes.
	Analyzing Errors	Identifying and articulating error in thinking
	Constructing Support	Building system of support for assertions
	Analyzing Perspectives	Identifying multiple perspectives on an issue and examining the reasons or logic behind each
Knowledge Utilization	Investigation	Identifying and resolving issues about which there are confusions or contradictions
	Problem Solving	Overcoming constraints or limiting conditions that are in the way of pursuing goals
	Experimental Inquiry	Generating and testing explanations of observed phenomena
	Invention	Developing unique products or processes that fulfill perceived need

2.2.2. Trial. The trial of the test was conducted at 6 schools in Bantul District. The location of the trial is based on the distribution of the National Examination in Bantul Regency. This study used 2 classes in each school with a total of 12 classes used, while the number of students who were the subjects of the trial were 322 students. The following details the number of test participants in table 2.

Table 2. Instrument testing subjects.

School	Sum of class	Sum of student
A	2	52
B	2	63
C	2	56
D	2	65
E	2	46
F	2	40
Total	12	322

2.2.3. *Assembling the test.* After the test is complete, the final step is to assemble the test into the CAT. CAT-MARZANO is then ready to be used as a media for physics-based learning assessment websites with reasoning multiple choice as a form of higher order thinking skills physics of Marzano's version. Home page of CAT-MARZANO in Figure 1 and home page for admin in figure 2.

**Figure 1.** Home page of CAT-MARZANO.**Figure 2.** Home page of CAT-MARZANO for admin.

3. Results and Discussion

The final product that has been made in this study is a website-based test with computer media. This test contains physics questions that were developed according to Marzano's taxonomy. The form of questions developed was in the form of reasoning multiple choice questions with as many as 136 items with 8 anchors in 4 question packages. The following discussion and results obtained from this study.

3.1. Test validation

Based on the results of validation by several experts, the Physics test with the HOTS version of Marzano was declared valid and feasible for use with aiken V values in the range 0.74 - 1.00 for all items.

3.2. Test estimation

Quest analysis is used to determine the estimation results of the test instrument by looking at the outputs on the Quest program. A summary of the estimation results can be seen in Table 3.

Table 3. Estimation results of test instrument.

No	Detail	Item Estimation	Test Estimation
1	Average and Standard Deviation	-0.01 ± 0.35	-0.61 ± 0.27
2	Average and Standard Deviation INFIT MNSQ	1.00 ± 0.07	0.99 ± 0.14
3	Average and Standard Deviation OUTFIT MNSQ	1.00 ± 0.09	1.00 ± 0.18
4	Average and Standard Deviation INFIT t	0.02 ± 0.59	0.01 ± 0.79
5	Average and Standard Deviation OUTFIT t	0.04 ± 0.50	0.06 ± 0.66
6	Reliability	0.68	0.68

3.3. Goodness of Fit

Testing the application of goodness of fit for each item with an average value of INFIT MNSQ must be in the range of 0.77 to 1.30 and the average value of INFIT t is in the range of values of -2.0 to 2.0. Based on the analysis, the MNSQ INFIT value is obtained between 0.93 to 1.07, so it can be concluded that each item is 136 items fit against the PCM 1 PL model. Example of MNSQ INFIT diagram for physics HOTS of Marzano's version can be seen in figure 3.

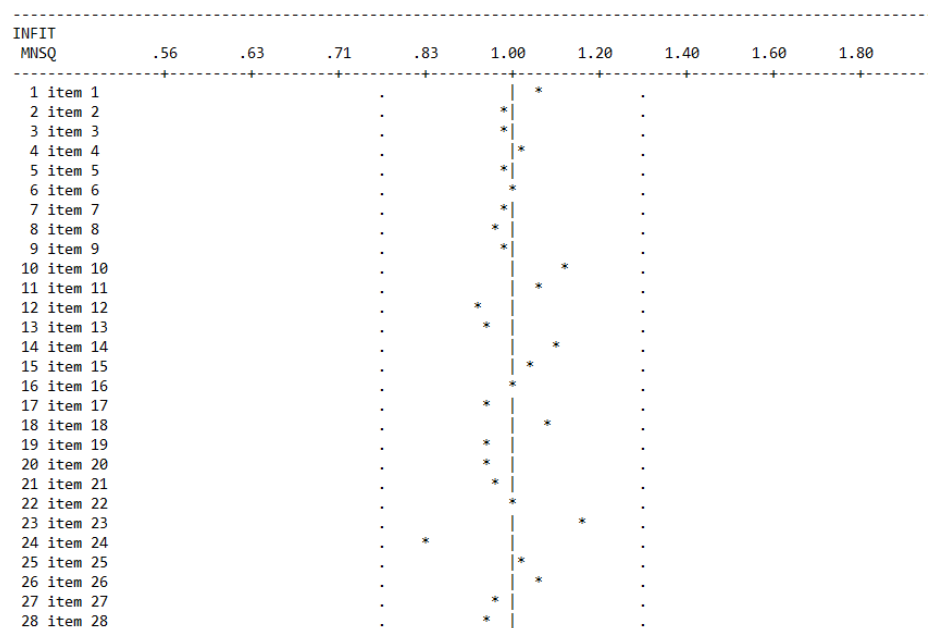


Figure 3. Example of MNSQ INFIT diagram for Physics HOTS of marzano'sversion.

3.4. The difficulty of the items

The results of the level of difficulty of each aspect and sub-aspect can be seen in table 4. The sub-aspect of investigation has the lowest level of difficulty (-0.30), while the sub-aspect of constructing support has the highest level of difficulty (0.23).

Table 4. Levels of difficulty particulars of aspects and sub aspects.

Aspect	Sub Aspect	Value of b	Category (%)			
			1	2	3	4
Knowledge recovery	Abstracting	0.10	54.35	19.98	15.73	9.94
	Decisions making	-0.20	46.48	21.95	17.60	13.98
Understanding	Inductive reasoning	0.03	54.14	15.01	18.94	11.90
	Deductive reasoning	0.21	58.80	18.32	15.11	7.76
Analysis	Comparing	-0.25	45.13	20.19	14.49	20.19
	Classifying	-0.04	51.45	16.67	18.01	13.87
	Analyzing errors	0.20	58.80	18.12	14.49	8.59
	Constructing support	0.23	59.01	18.84	13.87	8.28
	Analyzing perspectives	0.18	59.21	18.63	14.91	7.25
Knowledge utilization	Investigation	-0.30	45.03	13.77	17.70	23.50
	Problem solving	-0.05	51.45	20.70	15.84	12.01
	Experimental inquiry	0.04	57.84	13.66	16.69	11.80
	Invention	-0.04	50.62	17.49	20.08	11.80

3.5. Reliability

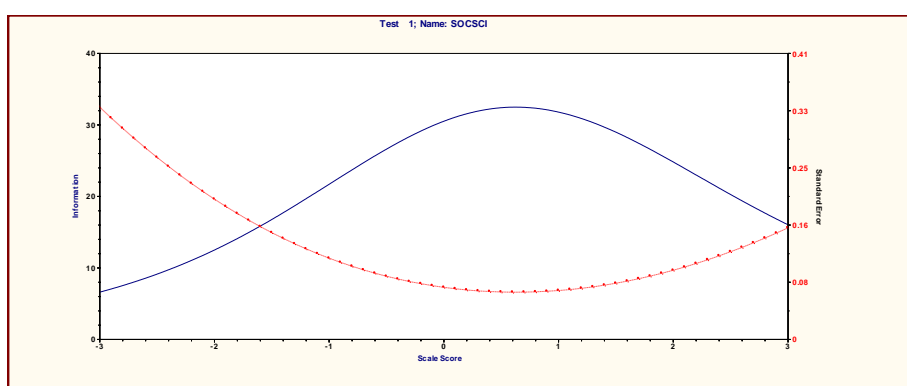
In addition to testing compatibility. The output of the Quest program also displays the estimated reliability of the test instrument set. Based on the results of the analysis with the Quest program with a Summary of item Estimates value of 0.68. From the figures obtained so that it can be concluded that the reliability of the test instruments can be categorized as sufficient.

3.6. Item Characteristic Curve

Item Characteristic Curves (ICC) and difficulty index of item test display item characteristics. Based on the analysis using the Parscale program, there were 136 item characteristic curve results. Figure 4 shows an example of ICC for item 1.

3.7. Information Function and Standard Error of Measurement (SEM)

Based on the analysis by Parscale obtained information functions and Standard Error of Measurement (SEM) like Figure 4. that the test is suitable for students with higher-order thinking skills in Physics (θ) in the high category, namely $-1.6 \leq \theta \leq 3.0$.

**Figure 4.** Relationship of information functions and SEM throughout all items

4. Conclusion

This paper has presented the development of CAT-MARZANO as an assessment media in the Industrial Revolution 4.0. The result of this research ascertained that CAT-MARZANO could be used as a media of assessment in response to the challenges of the industrial revolution 4.0. CAT-MARZANO is

declared valid and reliable for use in physics learning assessments. CAT-MARZANO is valid based on Aiken's V index in the range 0.74-1.00 and is effectively used with an effectiveness of 82.80%.

References

- [1] Schunk D H 2012 *Learning Theories an Educational Perspective (6rd ed.)* (Boston: Pearson Education)
- [2] Stiggins R J and Chappuis J 2012 *An Introduction to Student Involved Assessment for Learning (6rd ed.)* (Boston: Pearson)
- [3] Ju'subaidi 2017 *Conf. Proc. of Int. Conf. on Research in Education (ICRE)* 287-311
- [4] Bayrak B K 2013 *Mevlana Int. J. of Education (MIJE)* **3** 19-2610.13054/mije.13.21.3.2
- [5] Winarti. Carl. Suparmi. Sunarno W. and Istiyono E 2017 *Journal of Physics: Conference Series* **795** 1-510.1088/1742-6596/795/1/012052
- [6] Istiyono E. Mardapi D. and Suparno 2014 *Conf. Proc. of Int. Conf. on Educational Research and Evaluation (ICERE)* 101-10
- [7] Colwell N M 2013 *Journal of Education and Training Studies* **1** 50-6010.11114/jets.v1i2.101
- [8] Nikou S A and Economides A A 2013 *Conference: 6th Balkan Conference on Informatics: Thessaloniki* 107-14
- [9] Lu P and Cong X 2016 *J. of Phys.: Conf. Ser.* **710** 1-10 10.1088/1742-6596/710/1/012029
- [10] Giesinger J M, Kuster M S, Holzner B, and Giesinger K 2013 *The J. of Arthroplasty* **28** 418-2210.1016/j.arth.2012.08.026
- [11] Muthmainnah and Istiyono E 2019 *J. of Phys: Conference Series* **1233**: 1-910.1088/1742-6596/1233/1/012037
- [12] Oriondo L L and Antonio E M D 1998 *Evaluating Educational Outcomes (Test. Measurement. and Evaluation)* (Florentino St: Rex Printing Company)
- [13] Megawan M and Istiyono E 2019 *J. of Phys: Conference Series* **1233**1-710.1088/1742-6596/1233/1/012068
- [14] Rezaie M and Golshan M 2015 *International Journal of Educational Investigations* **2**128-37
- [15] Larson J W and Madsen H S 2011 *Calico J.* **2** 32-4
- [16] Ozyurt H, Ozyurt O, Baki A and Guven B 2012 *J. Expert System with Application* **39** 9837 – 4710.1016/j.eswa.2012.02.168
- [17] Larson J W and Madsen H S 2011 *Calico J.* **2** 32-4
- [18] Ozyurt H, Ozyurt O, Baki A and Guven B 2012 *J. Expert System with Application* **39** 9837 – 4710.1016/j.eswa.2012.02.168
- [19] Thompson N A and Weiss D J 2011 *Practical Assessment Research & Evaluation* p 16
- [20] Jamil M, Tariq R H and Shami P A 2012 *TOJET: The Turkish Online J. of Educational Technology* **11** 371-81
- [21] Marzano R J, Pickering D J, Arredondo D E, Blackburn G J, Brandt R S, Moffett C A, Paynter D E, Pollock J E and Whisler J S 1997 *Dimension of Learning Trainer's Manual 2nd Ed* (Colorado: McREL)