Computer-based Adaptive Test Development Using Fuzzy Item Response Theory to Estimate Student Ability

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

[1]Department of Educational Research and Evaluation, Yogyakarta State University, Indonesia
[2]Faculty of Science and Technology, State Islamic University of Sultan Syarif Kasim Riau, Indonesia

Received June 13, 2020; Revised July 23, 2020; Accepted August 25, 2020

Cite This Paper in the following Citation Styles

(a): *[1] Fitri Wulandari, Samsul Hadi, Haryanto , "Computer-based Adaptive Test Development Using Fuzzy Item Response Theory to Estimate Student Ability," Computer Science and Information Technology, Vol. 8, No. 3, pp. 66 - 73, 2020. DOI: 10.13189/csit.2020.080302.*

(b): *Fitri Wulandari, Samsul Hadi, Haryanto (2020). Computer-based Adaptive Test Development Using Fuzzy Item Response Theory to Estimate Student Ability. Computer Science and Information Technology, 8(3), 66 - 73. DOI: 10.13189/csit.2020.080302.*

Abstract The field of computing has developed so rapidly. Various theories of computational evolution to support human needs are continually being pursued; one of them is the field of education, especially in terms of teaching, testing, and evaluation of exam results. This study aims to develop computerized adaptive tests (CAT) to measure the student's abilities. Students will be measured for their cognitive abilities in Mathematics and Science subjects. It starts with developing a question bank that has been tested with 720 students to classify items based on its characteristic, i.e., easy, medium, and challenging. This research uses the item response theory approach with the model 2 logic parameters (2PL), namely item difficulty and item difference power. The selection of test items for each participant will depend on the response of the previous answer. Fuzzy algorithm is used in analyzing test items through four stages, namely fuzzification, implications, inference, and defuzzification. Meanwhile, to measure the ability of test-takers, the maximum likelihood estimation method, MLE, is used. Based on the testing of 73 students, it was found that each student received a different test item, both in the number of questions and the level of difficulty of the questions, according to student's abilities. The results of the CAT program's measurement of the test taker's ability estimation were stated to be more effective compared to conventional methods, as indicated by the average test length of 15 items compared to traditional tests, which had a length of 50 items. Therefore, the CAT program with the fuzzy item response theory can be used as support to measure students' abilities.

Keywords Computerized Adaptive Test, Maximum Likelihood Estimation, Fuzzy Algorithm

1. Introduction

With the development of the world and information technology, human behavior changes over time. This has also changed the development of the education system in the world and especially in Indonesia. The enhancement of the education system can be seen from the change in the education system, which includes learning, teaching, curriculum, learning methods, learning tools, facilities, and infrastructure, as well as graduate competencies from time to time. The current curriculum in the education system in Indonesia not only emphasizes the achievement of quantitative objectives in the form of test scores for several academic subjects but also emphasizes process-based assessment and student achievement [1]. Students are given more opportunities to choose the subjects they are interested in, to learn and develop their potential more flexibly based on their general basic skills (intelligence), talents, interests, and personality characteristics.

High school students in Indonesia must determine specialization or majors. There are three specialization groups, namely science, social, and language. Efforts to determine student specialization can be done by conducting cognitive tests on subjects that support. A test according to Mardapi [2] can determine learning achievements or competencies that have been achieved by students. Test results are information about the characteristics of a person or group of people in terms of their cognitive abilities or skills. Anderson and Kratwohl [3] show that cognitive domain learning is oriented towards thinking abilities, including simpler skills and problems solving skills. This testing activity is one way to predict the ability level of students indirectly, in response to a number of stimuli or questions. The test results are expected to produce data with as few errors as possible. Therefore, to get accurate data needed valid and reliable tests.

The use of computer technology to improve the quality of test results has been widely carried out. In the era of advanced technology and information, it is very feasible to conduct computer-based tests [4]. Reference [5-7] show that testing using a computer is not only able to produce tests that are fast and accurate but can also increase the effectiveness and efficiency of test implementation and maintenance. With advances in technology, paper, and pencil-based testing (PPT) has decreased due to the length of time in the management of tests and feedback [8]. In addition, the use of computers as a test medium has been widely developed, for example, in testing the English Language [9,10]. According to [11-13] in addition to testing, in the process of learning, teaching, and assessing, the use of computers has occupied a fairly comprehensive scope.

The use of computers has occupied a wide scope, recently computerized adaptive test (CAT) has been developed in various aspects. CAT is an adaptive based media where test-takers accept test questions according to their abilities[14]. Reference [13] further explains that CAT-based testing can improve efficiency and accuracy as well as practicality in its implementation [16]. CAT also optimizes managed items and can produce the most significant information in measuring the ability of test-takers [17]. CAT generally requires fewer items than long-form instruments and can achieve the same precision [16].

The use of CAT in each test has the aim to utilize the IRT invariant property in creating an algorithm, i.e., each test taker will receive test items that have been adjusted to the ability of individual test-takers. Hence, the questions given are not questions that are too difficult or too easy for individuals to test takers [18]. The same thing also stated by [19] that for accurate measurements, the level of difficulty of psychological tests must be in accordance with the ability of test-takers. A test will provide the most appropriate measurement of the test taker's abilities when the level of difficulty of the test is adjusted according to the test taker's ability level.

One important thing that needs to be considered in the preparation and development of adaptive tests is the procedure of analysis and selection of items. This is an important part because the quality of the test instrument is also determined by the quality of the items in it. In the selection of test items, Leung et al. [20] suggested that item selection control, minimizing test overlap, and efficient use of item groups are some of the important issues in designing computerized adaptive testing (CAT). Item selection is intended to enhance the accuracy of the test material [21]. Therefore, we need clear parameters about the characteristics of the material to be given.

In this study, CAT adaptive tests were developed using fuzzy algorithms in the item selection mechanism. The input for this algorithm is item difficulty, item difference power, and test the participant's response. The resulting output is the certainty of the selection of test items that have different power and difficulty level of items following the test participants' responses. This output is generated through a fuzzy inference mechanism in the form of further test items that will be given to test participants. The developed test was aimed to measure the ability of test participants in the specialization of science, so the subjects tested were mathematics and science abilities.

2. Methods

This study used a Research and Development (R&D) approach. There are three stages in the development of CAT, namely (1) building a question bank, (2) Selecting test items using Fuzzy CAT, and (3) estimating student ability.

2.1. Building an Item Bank

The item bank is a system that contains a collection of test items with a specific purpose, including its utilization system [22]. In developing the item bank, the items were compiled into a test kit and then tested. A test set is said to be good if it has good item characteristics, so it must first be analyzed the characteristics of the item. This can be done with both the classical and modern approaches (item response theory) [23-32]. But along with the development of science and technology, the use of item response theory became increasingly popular. There are three assumptions underlying the item response theory, namely unidimensional, local independence, and parameter invariance. In item response theory, the relationship between the probability of answering correctly on a capability scale is expressed by a relationship with the item parameters used. The number of item parameters used determines the equation model.

2.2. Selection of Test Items using Fuzzy CAT

According to Wainer [33,34], adaptive testing is a test conducted for test-takers with the questions/items determined based on the participant's initial answers/responses. Computerized adaptive testing not only can efficiently shorten test time and reduce the number of test items but also can accurately estimate test takers' abilities[35]

In CAT, the computer is set to select and provide items, and then the computer will calculate and score the test participants' answers. The items given to the test takers are items adjusted to the test taker's response to the previous items. If the item is answered correctly, the next item is presented with a higher difficulty level. If the items are answered incorrectly, then items will be presented with lower levels of difficulty [36]

By using fuzzy algorithms, item selection becomes somewhat different. Fuzzy logic has so far used two ways to represent uncertainty, ambiguity, and degree of fuzzy[37] : words and membership functions. The use of words such as high, medium, and low, is used by Shabaninia [38] to determine the vagueness associated with variables. While the membership function provides a more mathematical view of fuzziness by setting membership values for each value the fuzzy variable can take. This study uses fuzzy variables, namely the different items, the difficulty level of items, and the response of test participants' answers. These parameters are processed through the membership function in the fuzzy set. The output obtained is the certainty of the test items that have differentiator power and the difficulty level of the items up or down depending on the response of the test takers. The output is done by an inference mechanism based on a fuzzy algorithm in the form of further test items that will be given to test participants.

The inference system, which is also known as fuzzy control, is a mechanism in fuzzy logic to determine decisions. The inference model used in this study is Tsukamoto[39]. Fuzzy logic algorithm to produce output is done through four stages, namely:

a. Fuzzification. Input variables and output variables are divided into one or more fuzzy sets, which are performed based on the selected membership function.
b. Implications, namely the formation of rules (rules), based on a knowledge base. According to the Tsukamoto method, the implication function used is min (smallest value)
c. Inference, the affirmation of decisions based on the composition of the rules (rule base), is a collection of rules that are used as a basis for inference.
d. Defuzzification, i.e., confirmation of the results of inference based on a weighted average value

The input of the defuzzification process is the fuzzy set obtained from the inference mechanism for the composition of fuzzy rules. The output generated from this defuzzification process is a number in the fuzzy set domain. If a fuzzy set is given in a certain area, then a certain crispy value can be taken as the output of the defuzzification process.

The steps for selecting test items with fuzzy logic are shown in Figure 1:

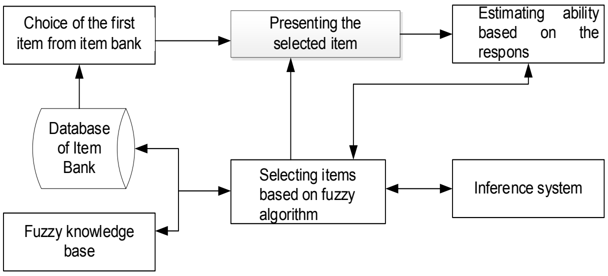

Choice of the first item from item bank
Presenting the selected item
Estimating ability based on the respons
Database of Item Bank
Selecting items based on fuzzy algorithm
Inference system
Fuzzy knowledge base

Figure 1. Selection of CAT Points with Fuzzy Logic Algorithms

Based on Figure 1, the process of selecting CAT items with fuzzy logic algorithm starts by selecting the first item from the question bank. After the items are selected, then the items are given to the test takers. Participants respond (right or wrong) to the item, then the level of the participant's ability is estimated again. Based on the input of item difficulty, item differentiator power, and the response of test participants' answers, the parameters are processed through the fuzzy logic function. The output is done by an inference system mechanism based on a fuzzy algorithm in the form of further test items that will be given to the test participants. This process continues and is terminated after as many items as specified have been given or after a precise estimation of the ability level or desired standard error measurement has been achieved.

2.3. Estimating the Ability of Students

Item response theory (IRT) is a psychometric theory that provides a basis for measuring the scale of test participants and items based on responses given to those items. Modern testing models with IRT are distinguished by the number of parameters, namely, one logistic parameter model (1 PL or Rasch model), two logistic parameters (2 PL), and three logistic parameters (3 PL)[31]. Reference [40] states that these parameters are item difficulty, item differentiator power, and guesses. This study uses the IRT model for the item dichotomy of two logic parameters (2PL), namely item difficulty, and item differentiator power, mathematically formulated as follows

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

Information :

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

θ : subject ability level (as a free variable)

a_i : index of differentiator power from item "i"

b_i : difficulty index of item "i"
e : natural number whose value is close to 2.718
n : number of items in the test

Three IRT concepts used in developing CAT are (1) item information function (IIF), (2) standard error measurement (SEM), and (3) capability level estimation. The item information function is stated with I_i (θ), which is a function that provides information by the item i on θ. Each item has information that is how well it can distinguish among test takers with the same ability at different levels of ability. Mathematically, the item information function fulfills the following equation

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

Information :
I_i(θ) : information function of item i
Pi(θ) : the opportunity for participants with the ability θ to answer item i correctly
Pi' (θ) : derivative of the function Pi (θ) with respect to θ
Qi(θ) : the opportunity for participants with the ability to answer item i incorrectly

Equations (1) and (2) show that the value of information depends only on the item parameters (a and b) and ability. Thus, for each level of ability (θ), the contribution of information for each item of question bank can be calculated. The test information function is the sum of the information functions of the test item compiler, reference [32] describing how accurately the test device estimates the level of different abilities. The greater the information at the given capability level, the more accurate the ability is estimated from the test set.

Standard Error of Measurement (SEM), which is the standard error of measurement, is closely related to the information function. The test information function is inversely quadratic with SEM, so the greater the test information function, the smaller the SEM or vice versa. The relationship between the two is stated by[31]:

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta}} \qquad (3)$$

The assessment of students' abilities is done first by calculating the value of pi (θ) and qi (θ) of each test item. The value of ability (θ) is taken in the range of -3 to +3 with step 0.2. Furthermore, with the known values of pi (θ) and qi (θ) and θ can be calculated the value of Likelihood L (U|θ) with the following equation:

$$\mathrm{L(U|\,\theta)} = \prod_{i=1}^{n} p_i{}^{n} q_i{}^{1-n} \qquad (4)$$

Based on the results of L (U | θ) for θ value of -3 to +3 with the estimated ability of the test taker is θ from the L (U | θ) the highest (maximum). So the estimated character of the test taker's abilities is determined by the formula:

$$\text{Estimated } \theta = \text{Maximum } [\mathrm{L\ (U \mid \theta)}] \qquad (5)$$

The estimation states that the probability of the test taker with the ability character (θ) answering the items max L (U | θ) x 100% is correct. On the other hand, the probability of a test taker with the ability (θ) to answer the item [1-max L (U | θ)] x 100% is wrong.

3. Results and Analysis

3.1. Item Bank

Calibration using the R program produced 160 mathematics items and 151 science items. A summary of the statistical parameters of items in the Empirical item bank is presented in Table 1 and Table 2 below.

Table 1. Statistics of Mathematics Item Bank

Parameter	Mean	Deviation Std.	Minimum	Maximum
Discriminant Index (a)	0,9350	0,3667	0,17	2,47
Difficulty Index (b)	-0,3696	0,8256	-2,19	4,01

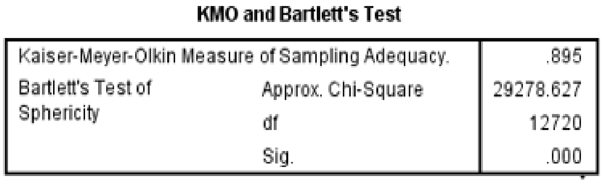
Table 2. Statistics of Science Item Banks

Parameter	Mean	Deviation Std	Minimum	Maximum
Discriminant Index (a)	0.8063	0.38703	0,14	2,22
Difficulty Index (b)	-0.318	1.16581	-3.88	3,86

An item is said to be good if the difficulty level (b) is from -2 to 2 [22], it can be concluded that each item has a normal difficulty level because it ranges from -2.19 to 4.01 for mathematical subjects, and between -3.88 to 3.86 for subject science. Likewise, the discriminant index value (a), produces good items from 0.17 to 2.47 for mathematical subjects and from 0.14 to 2.22 for Natural Sciences. As a reference, good items have a discriminant index from 0 to 2 [32].

One assumption can be proven one of them by using factor analysis to see the eigenvalues in the inter-grain covariance variant matrix. Data analysis with factor analysis was preceded by an analysis of sample adequacy. In this research, it has proven the unidimensional assumptions in the test participant data on Mathematics and Natural Sciences. The test kit was tested using a Computer Based Test (CBT) to 770 junior high school students in the city of Yogyakarta.

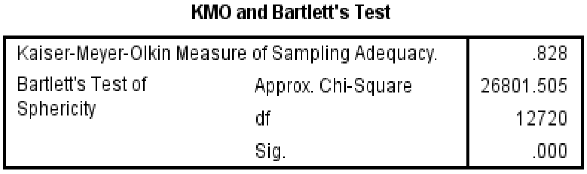
Based on the analysis of the adequacy of the sample in the mathematics test, the Chi-square value of the Bartlet test is 29278,627 with 12720 degrees of freedom. The p-value is less than 0.01 (see figure 2), while for the sciences subjects, the Chi-square value obtained in the Bartlet test amounted to 26801.505 with degrees of freedom 12720 and p-value less than 0.01 (see figure 3). These results indicate that the sample size used in this study was sufficient [41,42].



KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.895
Bartlett's Test of Sphericity	Approx. Chi-Square	29278.627
	df	12720
	Sig.	.000

Figure 2. KMO and Bartlett's test of Mathematics Test



KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.828
Bartlett's Test of Sphericity	Approx. Chi-Square	26801.505
	df	12720
	Sig.	.000

Figure 3. KMO and Bartlett's test of Science Test

Furthermore, the test participant's response pattern data was calibrated with the R program. Items that meet the ideal criteria are those that have different power levels in the range 0 to 2, and the difficulty level of -4.0 to +4.0 are selected as items on question bank for CAT purposes.

Table 3. Range of Mathematics Test Items

Group	Frequency	Percentage
• Ease	32	20%
• Moderate	95	59,4%
• Difficult	33	20,6%

Table 4. Range of Science Test Items

Group	Frequency	Percentage
• Ease	41	27,1%
• Moderate	80	53 %
• Difficult	30	19,9%

The number of items in the item bank is classified into three groups based on the difficulty level of the items, namely the difficult, medium, and easy item groups. The results of the item classification are presented in Tables 3 and 4.

The distribution of the group is quite good, with the portion of the medium difficulty level group is higher than the item groups with the difficulty level easy and difficult.

3.2. CAT Test Results

The CAT program has been tested with students. Testing by students is also called beta testing to see the performance of the program in displaying items in accordance with student responses, displaying students' abilities, and providing recommendations to students. The application trial was followed by 73 students. The following are the results of the science test from the students one based on the answer history. The results of the science test for students1 in the form of a history of answers and abilities (θ) are presented in table 5. The number of science questions received by Student1 is 17 items. Response answer 1 states that the answer is correct, and response answer 0 means the answer is wrong.

Table 5. History of Science Test Answers

No.	Code	Difficulty Level	Different Power	Response	SEM	FIB	Theta
1	FIS2033	-0.011	0.549	1	46,300	0,0466	3
2	FIS1008	0.694	0.908	1	30,050	0,1107	3
3	FIS2041	0.913	0.922	0	0,8605	13,507	1
4	BIO2057	0.553	0.227	1	0,9588	10,877	1.5
5	FIS2053	0.672	1.047	1	10,255	0,9509	2
6	FIS1011	1.018	0.772	1	11,883	0,7082	2.5
7	FIS2057	1.035	0.899	0	0,6357	24,744	1.5
8	BIO1028	0.47	0.572	1	0,7584	17,387	2
9	FIS2037	0.772	1.047	0	0,4951	40,794	1
10	FIS1009	0.448	0.532	1	0,5501	33,049	1.5
11	KIM012	0.661	0.195	0	0,4823	42,990	1
12	BIO1016	0.442	0.346	1	0,5415	34,105	1.5
13	FIS2031	0.755	0.462	0	0,4696	45,347	1
14	FIS2055	0.385	0.9	1	0,5094	38,534	1.5
15	FIS1007	0.719	0.807	1	0,4887	41,867	1.5
16	FIS1001	1.202	0.573	1	0,4757	44,189	1.5
17	KIM030	0.639	0.429	0	0,4693	45,396	1.5

Student's 1 answer history table contains information about the question number, question code, level of difficulty, power difference, response answers, SEM, FIB, and theta. In the first point, the FIS2033 question code has a difficulty level of -0.011, a power difference of 0.549, and the response of the answer is correct. Fuzzy algorithm is used in determining the next test item, and the second test item is selected item FIS1008 with a difficulty level of 0.694. The selection of the second test item is following the rules used, i.e., if the response of the answer is correct, then the level of difficulty of the item is raised. Then the second test item was responded with correct student answers, and the level of difficulty of the third test was raised again, and the third test item was chosen with the FIS2041 question code and difficulty level 0.913. The third item gets the wrong answer so that in the fourth item, the difficulty level is reduced. The selection of the fourth item is following the rules and the items selected with the BIO2057 item code and the level of difficulty item 0.553 and so on until the 17th test item.

The number of items done by each test participant is different, which depends on the achievement of the stopping rule. The stopping rule is the difference between SEM ≤ 0.01 or the maximum number of items, which is 20 items has been reached. In the case of the Science test, the 17th item has reached a difference of SEM ≤ 0.01 so that the test item is stopped.

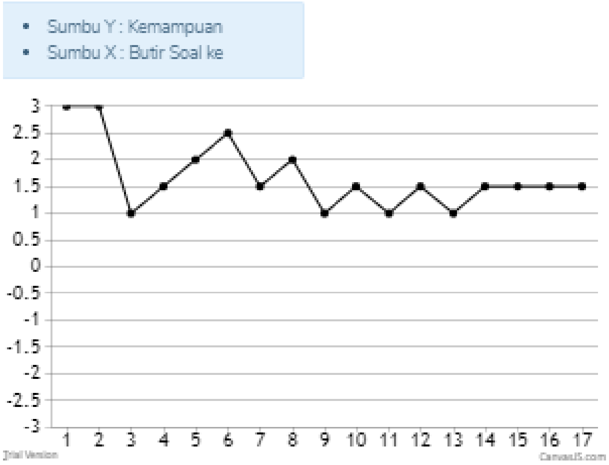


Figure 4. Estimated Student's 1 Ability Chart

The difficulty level up and down of the test item shows the ability of students (see figure 4). Start from the 14th item, there has been a stability of the ability, and the 17th item has reached the difference of SEM, which was 0.01. Out of the 17 items received by students 1, the responses of students' answers were mostly correct, and the level of difficulty of the test items received varied. In item 17, theta (θ) 1.5 was obtained, and after being converted using the scale of 0-100, the level of cognitive ability of science was 75.

Table 6. Cognitive Abilities of Student's 1 Science

Capability θ	L(U|θ)	L(θ|U)
-3	0.00000	0.0000
-2.5	0.00000	0.0000
-2	0.00000	0.0001
-1.5	0.00000	0.0007
-1	0.00001	0.0041
-0.5	0.00004	0.0184
0	0.00014	0.0613
0.5	0.00032	0.1438
1	0.00052	0.2293
1.5	0.00056	0.2482
2	0.00042	0.1882
2.5	0.00024	0.1058
3	0.00011	0.0469
Total	0.00225	1.0000

Based on table 6, it can be seen that the maximum value of L (θ | U) was 0.2482, with a position of ability (θ) of 1.5, illustrating that the results of the student's Science ability test were 1.5. This implies that the opportunity for students with the ability [θ] = 1.5 to answer the test items correctly was 24.82%. If table 6 graphs the likelihood function, it will look like in Figure 5.

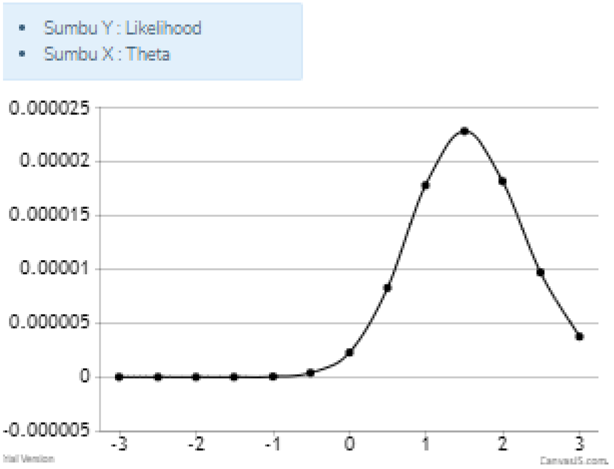


Figure 5. Graph of Likelihood Function of Student's 1 Item Response

3.3. Estimated Ability (Theta)

In addition to knowing the ability of the system to display items in accordance with the ability of students, beta testing was also used to determine the strength of the system to predictabilities and provide specialization recommendations based on the results of the estimated ability. Beta test results tested on 73 students are presented in table 7 below.

Table 7. Test results for the CAT Program

Subjects	Item Count			Score		
	Maximum	Minimum	Average	High	Low	Average
Mathematics	20	10	15	100	33	77.97
IPA	20	11	16	100	41.67	77.73

Student test results for each subject package were grouped into three, namely low group (L) for value <60, medium group (M) for value ≥ 60 up to value <80, and the high group (H) for value ≥ 80. The amount and the percentage distribution of students' level of ability, as shown in table 8. If students get high scores in both subjects (2H), or one high and one moderate (HM or MH), or both moderate (2M), then it is recommended to take a science specialization. Based on the distribution of students' ability levels, 49 students were obtained to be approved in the science specialization.

Table 8. Distribution of ability levels

Subjects	Quantity	Student Ability		
		Low	Moderate	High
Mathematics	Frequency	14	23	36
	Percentage	19.18%	31.51%	49.32%
Science	Frequency	14	24	35
	Percentage	19.18%	32.88%	47.94%

4. Conclusions

Based on the results of testing the CAT program in this study, it can be concluded that the CAT program with the fuzzy response item theory model produces different items for each student based on their ability level. Besides, the level of difficulty items received by students in accordance with the characteristics of item information. This is consistent with the nature of CAT theory which demands adaptability in tests. The nature of adaptability is contained in the fuzzy theory inference system that can determine the decision that each student must receive the right number of items, and each student must receive the right test items according to their ability characteristics.

The CAT program measurement results on the test taker's ability estimation were stated to be more effective compared to conventional methods, as indicated by the average test length of 15 items compared to traditional tests, which had a length of 50 items. Therefore, the CAT program with the fuzzy item response theory can be used as support to measure students' abilities.

The CAT program in this study is also able to estimate students' abilities and recommend specialization. Recommendations for specialization in cognitive tests are only one part of the test that can be considered in recommending the specialization of students. Some other tests, such as psychological tests and aptitude tests, can be used as a complement to these recommendations.

The CAT program in this study is also able to estimate students' abilities and recommend specialization so that they can be used as a support to measure students' abilities and interests. Recommendations for specialization in cognitive tests are only one part of the test that can be considered in recommending the specialization of students. Some other tests, such as psychological tests and aptitude tests, can be used as a complement to these recommendations.

Acknowledgments

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

REFERENCES

[1] S. Gultom, *Pedoman peminatan peserta didik*. Jakarta: Kementerian Pendidikan dan Kebudayaan, 2013.

[2] D. Mardapi, *Teknik penyusunan instrumen tes dan non-tes*. Yogyakarta: Mitra Cendikia Press, 2008.

[3] L. W. Anderson and D. R. Kratwohl, *A taxonomy for learning, teaching, and assessing*. New York: Addison Wesley Longman. Inc, 2011.

[4] H. Khoshsima and H. S. M. Toroujeni, "Computer Based Testing : Score Equivalence and Testing Administration Mode," vol. 12, no. 10, pp. 35–55, 2017.

[5] E. Georgiadou, Triantafillou, Evangelos, and A. A. Economides, "Evaluation parameters for computer-adaptive testing," *Br. J. Educ. Technol.*, vol. 37, no. 2, pp. 261–278, 2006.

[6] T. S. Chee and A. F. L. Wong, *Teaching and learning with technology*. Singapore: Prentice Hall, 2003.

[7] P. A. Towndrow and M. Vallence, *Using IT in the Language Classroom : A Guide for Teachers and Students in Asia*, 3rd ed. Singapore: Longman Pearson Education South Asia Pte.Ltd, 2004.

[8] J. Boo and W. Vispoel, "Computer versus Paper-and-Pencil Assessment of Educational Development: A Comparison of Psychometric Features and Examinee Preferences 1," vol. 1, no. V, pp. 443–460, 2012.

[9] J. Jamieson, "Trends in computer-based second language assessment," *J. Appl. Linguist.*, vol. 25, pp. 228–242, 2009.

[10] A. Alwi, M. Mehat, and N. I. Arshad, "E-Semai Teaching Portal (ESTP): A Preliminary Study in Assisting the Teaching of Bahasa Semai," vol. 8, no. 1, 2016.

[11] R. E. Bennett, "Inexorable and inevitable: The continuing story of technology and assessment. The Journal of Technology, Learning and Assessment," *J. Technol. Learn. Assessment,* vol. 1, no. 1, pp. 1–23, 2012.

[12] M. Pommerich and E. M. Russell, "Developing Computerized Versions of Paper-and-Pencil Tests : Mode Effects for Passage-Based Tests," vol. 2, no. 6, pp. 3–44, 2004.

[13] H. M. Fadzil, "Designing infographics for the educational technology course: Perspectives of pre-service science teacher," *J. Balt. Sci. Educ.*, vol. 7, no. 1, pp. 8–18, 2018.

[14] N. A. Thompson and D. J. Weiss, "A Framework for the Development of Computerized Adaptive Tests," *Pract. Assesment Res. Eval.*, vol. 16, no. 1, pp. 44–65, 2011.

[15] E. Triantafillou, E. Georgiadou, and A. A. Economides, "CAT-MD : Computerized Adaptive Testing on Mobile Devices," *Int. J. Web-Based Learn. Teach. Technol.*, vol. 3, no. 1, pp. 13–20, 2008.

[16] D. Cella and Æ. R. Gershon, "The future of outcomes measurement : item banking , tailored short-forms , and computerized adaptive assessment," *Qual. Life Res.*, vol. 16, no. 1, pp. 133–141, 2007.

[17] S. M. Haley, W. J. Coster, and H. M. Dumas, "Accuracy and precision of the Pediatric Evaluation of Disability Inventory computer-adaptive tests (PEDI-CAT)," *Dev. Med. Child Neurol.*, vol. 53, no. 12, pp. 1100–1106, 2011.

[18] S. E. Embretson and S. P. Reise, *Item Response Theory*, 1st ed. New York: L. Erlbaum Associates, 2000.

[19] F. M. Lord, *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum, 1980.

[20] C. Leung, H.-H. Chang, and K. Tai Hau, "Item Selection in Computerized Adaptive Testing : Improving the a -Stratified Design With the Sympson-Hetter Algorithm," *Psychol. Meas.*, vol. 26, no. 4, pp. 376–392, 2002.

[21] L. L. Davis and B. G. Dodd, "Item Exposure Constraints for Testlets in the Verbal Reasoning Section of the MCAT," *Appl. Psychol. Meas.*, vol. 27, no. 5, pp. 335–356, 2003.

[22] H. Retnawati, *Teori Respon Butir dan Penerapannya*, 1st ed. Yogyakarta: Parama Publishing, 2014.

[23] P. Awopeju, O. A., PhD., Afolabi, E. R. I., "Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination," vol. 12, no. 28, pp. 263–284, 2016.

[24] M. Zoghi, V. Valipour, and A. Branch, "A Comparative Study of Classical Test Theory and Item Response Theory in Estimating Test Item Parameters in a Linguistics Test," *Indian J. Fundam. Appl. Life Sci.*, vol. 4, pp. 424–435, 2014.

[25] D. J. Ratnaningsih and Isfarudi, "Analisis butir tes objektif ujian akhir semester mahasiswa universitas terbuka berdasarkan teori tes modern," *J. Pendidik. Terbuka dan Jarak Jauh*, vol. 14, no. 2, pp. 98–109, 2013.

[26] N. Abedalaziz and C. H. Leng, "The Relationship between CTT and IRT Approaches in Analyzing Item Characteristics," *Malaysian Online J. Educ. Sci.*, vol. 1, no. 1, pp. 64–70, 2013.

[27] G. L. Thorpe, G. L. Thorpe, and A. Favia, "Data Analysis Using Item Response Theory Methodology : An Introduction to Selected Programs and Applications .," *Psychol. Fac. Scholarsh.*, pp. 1–33, 2012.

[28] G. Margono *et al.*, *Pengembangan Instrumen Penelitian Pendidikan*, 1st ed. Yogyakarta: Graha Ilmu, 2013.

[29] C. DeMars, *Peranan Asesmen dan Ujian Dalam Peningkatan Mutu Pendidikan nasoinal*. New York: Oxford University Press, Inc, 2010.

[30] l Crocker and J. Algina, *Introduction to classical and modern test theory*. New York: Holt, Reinhart, and Winston, Inc, 2008.

[31] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, "Fundamentals of Item Response Theory.," *Contemporary Sociology*, vol. 21, no. 2. p. 289, 1991.

[32] R. K Hambleton and H. Swaminathan, *Items response theory: principles and application*, 1st ed. Boston: Springer Netherlands, 1985.

[33] H. Wainer, *Computerized adaptive testing_ Aprimer*. New Jersey: Lawrence Erlbaum Associates, 1990.

[34] H. Wainer, E. T. Bradlow, X. Wang, and S. Johnson, *Testlet Response Theory and Its Applications*. Cambridge: Cambridge University Press, 2007.

[35] D. J. Weiss and D. J. Weiss, "Counseling and Education," vol. 37, no. 2, pp. 70–84, 2004.

[36] Mardapi Djemari, S. Hadi, and Haryanto, "Pengujian Belajar dan Penilaian Pendidikan Berbantuan Komputer," *J. Kependidikan Univ. Negeri Yogyakarta*, vol. 42, no. 2, pp. 130–143, 2012.

[37] F. Shabaninia, "Z-mouse : A New Tool in Fuzzy Logic Theory," *World J. Comput. Appl. Technol.*, vol. 2, no. 1, pp. 22–27, 2014.

[38] L. A. Zadeh, "Is there a need for fuzzy logic?," *Inf. Sci. (Ny).*, vol. 178, no. 13, pp. 2751–2779, 2008.

[39] J. Yan, M. Ryan, and J. Power, *Using fuzzy logic*. Englewood Cliffs: Prentice Hall Inc, 1994.

[40] V. Der Linden, J. Wim, and R. K. Hambleton, *Handbook of modern item response theory*. New York: Springer Verlag, 1997.

[41] A. A. Afifi and V. Clark, *Computer-Aided Multivariate Analysis*, 3rd ed. New York: Chapman And Hall/CRC, 1996.

[42] W. R. Dillon and M. Goldstein, *Multivariate Analysis Method and Application*. New York: John Wiley and Sons, 1984.