Journal of Theoretical and Applied Information Technology
28th February 2021. Vol.99. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMIZING OF ITEM SELECTION IN COMPUTERIZED ADAPTIVE TESTING BASED ON EFFICIENCY BALANCED INFORMATION

[1,2]JANU ARLINWIBOWO, [1]HERI RETNAWATI, [1]SAMSUL HADI, [1]BADRUN KARTOWAGIRAN, [3]GULZHAINA K. KASSYMOVA

[1]Postgraduate Program, Universitas Negeri Yogyakarta, Indonesia

[2]Faculty Science, Technology, and Mathematics, Universitas Muhammadiyah Kudus, Indonesia

[3]Abai Kazakh National Pedagogical University; Institute of Metallurgy and Ore Beneficiation, Satbayev University, Kazakhstan

E-mail: [1]januarlinwibowo.2018@student.uny.ac.id, [2]heri_retnawati@uny.ac.id, [3]samsul_hd@uny.ac.id , [4]kartowagiran@uny.ac.id, [5]g.kassymova@satbayev.university

ABSTRACT

Efficiency Balanced Information is an item selection method developed by Han (2009) with the aim of this research is to solve problems of other methods that have previously been widely applied in CAT. This study aims to explore the effectiveness of using Efficiency Balanced Information for item selection. This research will identify criteria that affect this method so that it is optimal when used in the CAT to predict student abilities. This research is simulation research with two simulation software, WinGen, and SimulCAT. Simulations are conducted by involving several variables, namely the ability estimation method (MLE, MAP, and EAP), item parameter estimation method (1PL, 2PL, and 3PL), Efficiency Balanced Information as an item selection method, and stopping rule based on SE which is 0.3. The simulation involves 150 items that characters have been defined in WinGen and the researcher run in SumulCAT 30 replications. Based on the results of the study it can be concluded (1) The most powerful item estimation method is 2PL, (2) the ability to estimate capabilities with EAP and MAP is relatively the same and the MLE indicates the ability interval that can be estimated narrower, (3) efficiency of administered items with EAP and MAP is relatively the same while MLE needs more items.

Keywords: *Efficiency Balanced Information, Item selection, CAT, Effectiveness, Ability Estimation*

1. INTRODUCTION

Tests are one important component in education. The test is used to determine the level of student understanding of knowledge and concepts [1]. [2] states that test results are the profiles used to define a student on a particular thing. [3] argue that test results can be used as data to reflect learning outcomes. [4] states that the test is a systematic procedure that can be used to identify how much has been obtained by students in the learning process. Thus the test is an instrument to see certain insights owned by students. In general, the results of a test can be used based on taking on policy. For students, the test results can be data for self-reflection, how much competence that was mastered in learning. As for the teacher, test results can be used as basic data for determining learning strategies. Benefits for schools, test result data can be used to assess the quality of learning performance. The government needs data on test results to map the quality of education.

A test result must be accurate, in the sense of being able to represent the students' ability clearly [4]. To be able to represent students' abilities, a test must pay attention to measurement errors. Errors are divided into two, namely random errors and systematic errors [5], [6]. Systematic errors are errors that are affected by the quality of the instrument, the measurement process, or a combination of both [7]. Thus a test must be based on a strong theory so that the instrument that is built can measure what should be measured [2] and built with the right procedure because this systematic error cannot be identified in a test result [7]. Random error is an unpredictable error. These errors can come from internal test takers or environmental influences [8]. This random error has a strong relationship with reliability. Therefore the reliability of a test must be proven and

guaranteed [9]. On a technical level, the test maker and organizer must be very careful. Test makers must ensure that the test design is good and the test organizer must be able to condition a test as the design. If a test result does not represent the student's condition then the use of the test result is not on target.

Efficiency and accuracy are the main issues in the test [10]. It is in this aspect that the CAT has advantages over traditional tests [11]. There are contradictions in traditional tests, namely the more test items (ineffective) the more accurate [12]. But keep in mind that there is an element of endurance test-takers who need attention. Usually in Indonesia, an agreement is taken on the number of items (40 for science and technology and 50 for social science) to produce an accurate test result but still within the reasonable limits of students' endurance in doing tests. It is important because students who actually master the material may become unable to work on problems within their abilities because they are already exhausted.

Regarding the substance of making a good test, [13] states that teachers must match the level of difficulty of the test with the level of student ability. Students should not deal with problems with a level of difficulty beyond their ability and also do not work on problems that are too easy [14]. If the questions in the test are beyond their ability, chances are they are giving up, guessing, or cheating, so the portrait of students' original abilities cannot be properly identified. If the questions are done too easily then the students' original ability also cannot be identified properly because students do not push to limits on their abilities. However, making test questions that are according to the ability of the students is not easy because the conditions of students are very heterogeneous.

[15] state that more clearly that a single test cannot measure accurately the ability of each student. A test that is compatible with students' abilities is needed. The solution that is [15] state is an adaptive test in which the test gives questions that are in accordance with students' abilities based on the item information. Thus, the test can be carried out involving fewer items but with a good standard of accuracy [16]. Giving items according to students' abilities makes the test results more accurate even though they only use a few items when compared to the classical test model. Each additional item increases accuracy (standard error decreases) [17]. The test requires more items only when students answer inconsistently based on their abilities (incorrectly answering questions with a lower difficulty level or being able to correctly answer questions with a very high difficulty level) [14]. This inconsistent pattern can also be indicated as an abnormal pattern in the process of running the test.

Such a system is very difficult to imagine or even impossible in a classic format. This concept becomes very possible in the era of rapid technological development. Technology makes all difficult things possible [18]–[20]. In the last few years, the assessment process has started to switch to computer formats, from offline to online systems. In Indonesia, national exams, college entrance selections, and various job tests have been carried out on a computer basis [18], [21]. Such a system makes the opportunity for implementing an adaptive test that combines a question bank with a particular question viewer algorithm very wide open [22]–[24].

In this era, the adaptive test that was developed by measurement experts was known as the Computerized Adaptive Test (CAT). The test utilizes computerized technology to support complexity in the choice of questions and infers students' abilities based on an algorithm that is in line with specific criteria. A question is chosen and administered by considering the track record (pattern) of students in answering or the ability of test participants when taking the exam [25]. Thus, the items taken for each student will vary so that the safety of the test can be more awake from cheating cases [26].

In the CAT system, the principle that is used is the selection of items test based on consideration of students' previous track record in answering questions of the test [27]. The question bank will make a unique arrangement of patterns for each examinee depending on each performance during the test [26], [28], [29]. Thus, the utilization of each item becomes more efficient and on target by remaining oriented to the level of measurement accuracy [30], [31].

The item selection method is a very crucial factor in CAT [10]. Each method expects the selection of items that can approach the students' ability in the process of working on the problem and show the minimum error standard. [32] developed a selection of items in CAT called

Efficiency Balanced Information. The item selection technique is designed to overcome the problem of two item selection methods that have been widely applied in the CAT such as the Fisher information method which were identified as not optimal at the initial item selection stage and the remaining items cannot function properly and secondly, the stratification method is judged to have a limited amount items for each stratum item available, c-parameters are not taken into account, and the method's ineffectiveness is related to variable length. According to [33], if the test is intended to estimate θ with a certain degree of precision, it is appropriate to retrieve items based on their information function.

The method developed by [32], named Efficiency Balanced Information (EBI), focuses on the efficiency items expected for item selection. Under this method, the selection of items will depend on the EBI index. The items with the largest EBI scores will be given to the test takers. Following is the formula for determining the EBI index.

$$EBI_i[\hat{\theta}_j] = \left(1 + \frac{1}{I_i[\theta_i^*]}\right) \int_{\hat{\theta}_j - 2\varepsilon_j}^{\hat{\theta}_j + 2\varepsilon_j} I_i[\theta_j]\, d\theta \quad (1)$$

There is a difference in θ_i^* in the items parameters estimation of 1PL, 2PL, and 3PL. θ_i^* is the same as b_i when the items parameters estimation used are 1PL or 2PL. Whereas when using 3PL for items parameters estimation, θ_i^* can be searched with the formula presented by [34], namely:

$$\theta_i^* = b_i + \frac{1}{Da_i} \ln\left(\frac{1 + \sqrt{1 + 8c_i}}{2}\right) \quad (2)$$

In the EBI procedure, the estimated ability level of participants (θ) will be adjusted to the level of difficulty of the items. The interval θ in the evaluation of item efficiency results from a standard estimation error (SEE; ε) and is set two times the SEE of $\hat{\theta}$ after the jth item is selected (notated $\hat{\theta}_j + 2\varepsilon_j$).

Statistical formulas always have a character for the subject being analyzed. Efficiency Balance Information is a development of previous statistical formulas that are used as a basis for determining selected items in CAT. The essence of the assessment process in CAT is the item response theory which involves various things such as parameter estimation methods, respondent abilities, information functions, and standard errors. From these aspects, it is divided into various techniques, for example, parameter estimation can use 1PL, 2PL, 3PL, or 4PL, and it is still developing now. Respondents' abilities were also estimated using various methods. In connection with the many aspects that have an influence on the analysis process, the study of the suitability of Efficiency Balance Information with other attributes is very important so that Efficiency Balance Information can be used appropriately to produce the best CAT performance [15], [28], [35].

Thus, researchers will conduct a search related to the effectiveness of the use of item selection based on Efficiency Balanced Information. The objective of the researcher in conducting this research is to explore the criteria of the method so that it is optimal when used as a basis for selecting items to predict students' abilities in CAT.

2. METHOD

This research is a simulation study that utilizes two simulation software namely WinGen and SimulCAT. WinGen is used to generate students' ability data and item parameters. SimulCAT is used to simulate CAT based on students' ability data and item parameters. The simulation is done by involving several variables as follows:

1. Ability estimation methods [36], namely Maximum Likelihood Estimation (MLE), Bayesian Maximum a Posteriori (MAP), and Bayesian Expected a Posteriori (EAP),
2. Item parameter estimation methods, namely 1-logistic parameter (1PL), 2-logistic parameter (2PL), and 3-logistical parameter (3PL),
3. efficiency Balanced Information as an item selection method, and
4. stopping rule based on SE is 0.3.

The process of generating data starts with generating students' ability data which will become student data participants in a CAT simulation. The provisions of the generation process carried out by the researcher are (~ N (0,1)) with 1000 students. Then the item data generation is done using the WinGen software with the following criteria.

Journal of Theoretical and Applied Information Technology
28th February 2021. Vol.99. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Description of Item Data Generation Attributes

Item Parameters	Distribution	1 PL		2 PL		3 PL	
		Min	Max	Min	Max	Min	Max
Discriminant (a)	Uniform	0.5	2	0.5	2	0.5	2
Difficulties (b)	Uniform			-3	3	-3	3
Pseudo guessing (c)	Uniform					0	0.2

Then generated 150 items of questions with predetermined characters and then analyzed with 1PL, 2PL, and 3PL. Person parameter and parameter items data are saved and then used as a CAT simulation database. Based on the database of parameter items and person parameters, CAT simulations with simulCAT (1PL, 2PL, and 3PL are simulated using MLE, MAP, and EAP ability estimation methods) each of which is replicated by 30 times. Examples are 1PL with MLE ability estimation that was conducted 30 replications, 1PL with MAP ability estimation that was conducted by 30 replications, 2PL with EAP ability estimation that was conducted by 30 replications, and so on.

The analysis was carried out in several stages as follows:

a. Comparing $\hat{\theta}$ and θ so that the correlation between the two is known. High correlation data indicates that the true score and observe score data are identical. In other words, the observed score can reflect the true score.
b. After the replication data has a high correlation between the true the ability and observe ability of students, an analysis is carried out on the average number of administrated items (the number of items worked on) by students until the CAT system stops. The analysis process was carried out in several groups which were a combination of the item parameter estimation method and the student's ability estimation method.
c. Based on CAT simulation data, an uncommon event is carried out, such as (1) the system does not stop until students have to work on an unreasonable number of questions, (2) the system stops immediately when starting to work, (3) the system stops with a value standard error, not below 0.3.
d. Data on unusual incidents will be recorded as the basis for determining the range of the system's ability to diagnose the respondent's ability. Based on these data, it can be concluded that EBI's performance in selecting items in CAT for various method combinations.

3. RESULT

The first step in this research is to analyze the relationship between the true abilities of students simulated using WinGen software and the results of the estimated ability from SimulCAT simulation. If the correlation is high (> 0.9) then the true score with the estimated score is near the same. The following is a comparison of true abilities and estimated ability that was conducted by 30 times replicated.

Table 2: Correlation of $\hat{\theta}$ to θ

Replication	Correlation of true ability with estimated ability								
	MLE			MAP			EAP		
	1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL
1	0.938	0.941	0.950	0.957	0.953	0.955	0.955	0.954	0.951
2	0.942	0.955	0.945	0.950	0.949	0.945	0.957	0.958	0.958
3	0.941	0.943	0.959	0.951	0.949	0.953	0.951	0.950	0.948
4	0.946	0.948	0.945	0.955	0.957	0.949	0.958	0.956	0.953
5	0.945	0.952	0.943	0.952	0.957	0.952	0.958	0.951	0.945
6	0.937	0.953	0.943	0.944	0.955	0.949	0.956	0.957	0.939
7	0.941	0.945	0.949	0.956	0.952	0.953	0.954	0.947	0.945

8	0.934	0.947	0.943	0.955	0.950	0.952	0.962	0.960	0.955
9	0.944	0.947	0.949	0.949	0.950	0.955	0.958	0.952	0.935
10	0.949	0.944	0.945	0.952	0.956	0.952	0.955	0.948	0.939
11	0.941	0.938	0.939	0.958	0.958	0.948	0.952	0.952	0.947
12	0.942	0.949	0.957	0.956	0.954	0.950	0.951	0.955	0.948
13	0.937	0.941	0.952	0.955	0.952	0.953	0.958	0.954	0.942
14	0.938	0.937	0.954	0.952	0.955	0.946	0.958	0.952	0.949
15	0.941	0.937	0.935	0.946	0.950	0.952	0.954	0.959	0.943
16	0.936	0.947	0.950	0.958	0.950	0.949	0.953	0.956	0.946
17	0.941	0.945	0.939	0.949	0.951	0.949	0.945	0.948	0.947
18	0.938	0.952	0.953	0.960	0.951	0.955	0.950	0.953	0.939
19	0.940	0.955	0.946	0.958	0.957	0.949	0.957	0.944	0.949
20	0.943	0.942	0.940	0.952	0.956	0.950	0.950	0.952	0.958
21	0.944	0.944	0.940	0.956	0.951	0.948	0.958	0.950	0.948
22	0.943	0.948	0.952	0.959	0.955	0.952	0.956	0.950	0.947
23	0.945	0.954	0.953	0.953	0.946	0.950	0.953	0.949	0.956
24	0.941	0.953	0.944	0.956	0.949	0.953	0.957	0.952	0.948
25	0.950	0.953	0.946	0.955	0.955	0.958	0.951	0.957	0.954
26	0.937	0.952	0.950	0.957	0.953	0.946	0.958	0.957	0.948
27	0.940	0.946	0.949	0.959	0.939	0.951	0.954	0.952	0.951
28	0.942	0.944	0.944	0.958	0.958	0.958	0.955	0.955	0.955
29	0.934	0.935	0.937	0.946	0.961	0.950	0.953	0.958	0.948
30	0.945	0.933	0.933	0.956	0.947	0.957	0.955	0.959	0.955

Based on the data in Table 2, it can be concluded that there is a very high relationship (> 0.9) in each simulation. Thus it can be concluded that there are similarities between the true ability parameters generated through WinGen as a database and the estimated ability parameters resulted from the simulation results through SimulCAT.

The results of CAT simulation with SimulCAT can be observed administered items data for each simulation with a combination of estimation ability methods and estimation items parameters involved. The following, in Table 3, is the average question data used (rounding off).

Table 3: Average Administered Questions Each Combination Techniques

Replication	1PL			2PL			3PL		
	MLE	MAP	EAP	MLE	MAP	EAP	MLE	MAP	EAP
1	19	18	18	10	8	8	11	9	9
2	19	18	18	10	8	8	11	9	9
3	19	18	18	10	8	8	11	9	9
4	19	18	18	10	8	8	11	9	9
5	19	17	18	10	8	8	12	9	9
6	19	18	18	10	8	8	11	9	9
7	19	18	18	10	8	8	11	9	9
8	19	18	18	10	8	8	11	9	9
9	19	17	18	10	8	8	11	9	9
10	19	18	18	10	8	8	11	9	9

11	19	17	18	10	8	8	11	9	9
12	19	18	17	10	8	8	11	9	9
13	19	18	18	10	8	8	11	9	9
14	19	17	18	10	8	8	11	9	9
15	19	17	18	10	8	8	11	9	9
16	19	18	18	10	8	8	11	9	9
17	19	17	18	10	8	8	11	9	9
18	19	18	18	10	8	8	11	9	9
19	19	18	18	10	8	8	11	9	9
20	19	17	18	10	8	8	12	9	9
21	19	18	18	10	8	8	12	9	9
22	19	18	18	10	8	8	12	9	9
23	19	17	18	10	8	8	11	9	9
24	19	18	18	10	8	8	11	9	9
25	19	18	18	10	8	8	12	9	9
26	19	18	18	10	8	8	11	9	9
27	19	18	18	10	8	8	11	9	9
28	19	18	18	10	8	8	11	9	9
29	19	17	18	10	8	8	11	9	9
30	19	17	18	10	8	8	11	9	9

Based on the definition in CAT, effectiveness is expressed in the number of items administered so MAP and EAP have better effectiveness compared to MLE. This is apparent in all item parameter estimation methods (1PL, 2 PL, and 3PL) where MAP and EAP require fewer items to stop the test compared to MLE.

However, from the simulation. It is indicated that there is some students' ability that cannot be identified properly with item selection techniques based on Efficiency Balanced Information. The following are a few examples of simulation results that show the inability of the CAT in estimating the ability of students assessed from the SE and the number of administered items.

```
301     1.036  18      1.140   0.292  100101010011100100      54,70,25,66,123,59,112,20,21,68,69,87,
302    -2.886  61     -3.170   0.290  0100000100000000100000000000000000010000000000000000000000001
303     1.717  18      1.563   0.297  101010100000101111      125,70,112,83,138,47,86,150,45,25,147,
797     0.508  18      0.116   0.299  100001011001000101      136,70,138,66,144,57
798    -3.576  0       0.251   9.999
799     0.554  18      0.475   0.297  100101000010111001      16,70,75,127,59,118,
```

Figure 1: Examples of CAT Response Patterns (student ID students' true ability, administered items, students' ability estimation, standard error (SE), response patterns, number code of administered item)

Based on the simulation data, there is an example of 2 unusual patterns. The first case is the second pattern (students with ID 302) from above which shows that there are a lot of administered items (61 items) for true abilities -2.886. The unusual case is apparent because the average of the model stops the personal test (SE has reached a value below 0.3) in question 18. Thus, the CAT by utilizing the item selection Efficiency Balanced Information has difficulty measuring that ability. It is assumed that the ability of -2.886 is too small to be observed. The second example is the fifth pattern from the top (students with ID 798 who have the true ability -3,576) showing that none of the items were administered and SE was in position 9999. This shows that the CAT is not moving and unable to measure the ability of students with these abilities. There is a conjecture similar to Example 1. The ability of -3.576 is too small to be observed in a CAT based on the efficiency balance

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

information. In 30 replications in each combination of models found many similar cases. Thus, the analysis must continue to map the simulation results to identify the ability intervals of students who can be analyzed with the CAT based on Efficiency Balanced Information.

Effectiveness is also traced based on the standard error for each pattern of student answers in the simulation. In the simulation set stopping rule based on the SE that is SE less than 0.3. Tracing the maximum student ability that can be detected so that the test can stop in accordance with the command (resulting in a standard error less than the specified stopping rule criteria). To be able to see carefully, the ability intervals were identified ranging from -3.5 to 3.5 with the distance between abilities 0.01 (identification of SE against abilities -3.5; -3.499; -3.498; -3.497; ...; 3.497; 3.498; 3.499; 3.5. Found several abilities that cannot be detected properly by CAT based on Efficiency Balanced Information because the simulation stopped but the SE did not touch the number 0.3. Based on the tracing data and item administered, it was found that the ability intervals of students which can be concluded by CAT based on Efficiency Balanced Information are are shown in Table 4.

Table 4: Interval Efficiency Balanced Information Table in Detecting Students' Abilities

Methods		Abilities	
		Lower Bound	Upper Bound
EAP	1PL	-2.81	3.17
	2PL	-2.72	3.5
	3PL	-2.45	2.91
MAP	1PL	-2.77	3.18
	2PL	-2.88	3.24
	3PL	-2.38	3.11
MLE	1PL	-2.52	3.01
	2PL	-2.65	2.97
	3PL	-2.33	2.84

Detections made in Table 4 are based on information items and the number of items administered. The data show that there are variations in each combination of the ability estimation method and item parameter estimation in 30 replications. The data shows the minimum and maximum abilities that can be estimated in CAT based on efficiency balance information with the usual number of items of administration and reach a standard error of 0.3 or less.

If seen from the item estimation method, 2PL has the longest interval in estimating students' abilities either through EAP, MAP, or MLE. When compared between methods of estimating students' ability, MLE is the method whose performance is the least because it has the shortest interval among other methods. Likewise, when looking at the lower and upper limits of the ability of students who are able to detect, the EAP and MAP performance is better than MLE with an indication that MLE has a lower limit that is upper than others and an upper limit that is lower than others. This means that MLE is unable to detect certain low and high abilities that can be detected by EAP and MAP.

4. DISCUSSION

Efficiency Balanced Information is an item selection system in CAT that utilizes the item information function. Simulation results show that there are differences in results that are influenced by the method of estimating students' ability. The Bayesian method (showing identical results between EAP and MAP) is more efficient than MLE when used in a CAT based on Efficiency Balanced Information.

The fundamental thing that distinguishes between MLE and Bayesian is Bayesian assumes a normally distributed population [37]. In Bayesian or also known as posterior Likelihood [38], the estimation results by the MLE method are multiplied by the values in the normal distribution according to certain ability intervals. Thus the distribution is closer to the mean [37]. Figure 2 is a

comparison of the ability estimation curve with the Likelihood method and the Likelihood method multiplied by the distribution in the normal distribution (Bayesian).

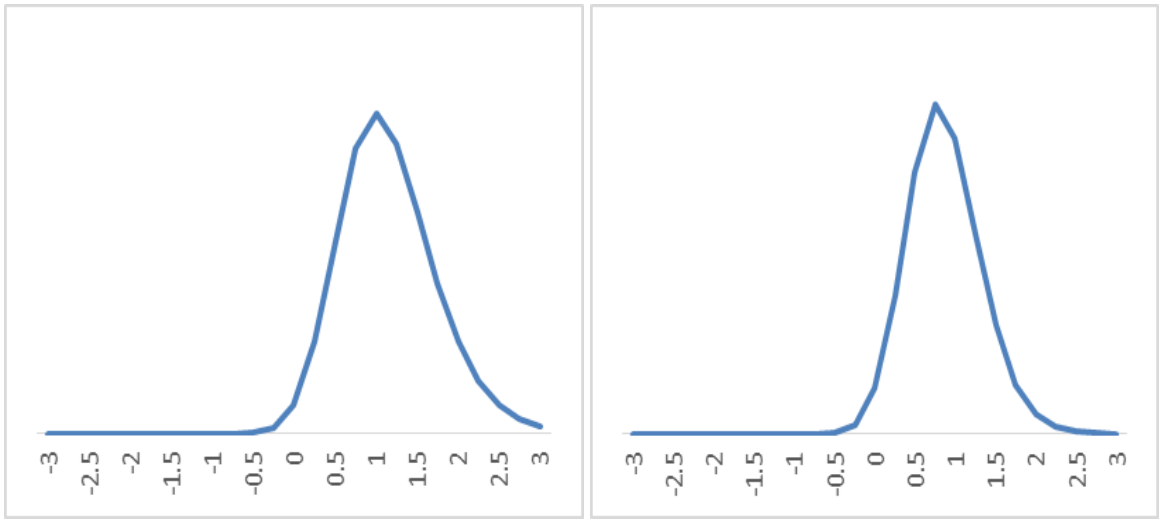

Figure 2: The Curve of Ability Estimation Using Likelihood Method (left) and The Curve of Likelihood Method Results Multiplied by Distribution in Normal Distribution (right)

Information about a test is obtained from the sum of the value of information items by utilizing probability data to answer correctly a student [15], [35], [39]. Thus it is reasonable if the value of item information is influenced by the students' ability estimation method. The next effect is to influence the stopping rule in the CAT because the criteria for stopping the test are a particular SE in this simulation, where SE according to [40] and [15] can be determined based on the information value of a test. Therefore, it seems clear that the assumption of a posteriori has an impact on the number of items administered.

The simulation results are in line with the research of [41] which states that the performance of EAP and MAP in estimating students' abilities is better than that of MLE. In that study, it was said that MLE has lower accuracy compared to the two other methods in estimating the ability of test-takers. Bayesian has a better ability to estimate ability [42] especially if the number of samples is small [43] and Bayesian has a better Mean Square Error (MSE) [44]. In the context of conventional tests, accurate is a small error (the smaller error, the more accurate), while in CAT can be translated as the efficiency of administered items. In CAT, the fewer items administered the more efficient the tests [45]. Ensuring quality education plays a key role in competency development in higher education [46].

5. CONCLUSION

Based on the results of the study it can be concluded that the implementation of CAT based on efficiency balance information is (1) the most powerful item parameter estimation method is 2PL, (2) performance of estimated students' abilities with EAP and MAP is relatively the same while MLE is indicated to have an ability interval that can be estimated narrower, (3) related to the number of items administered, the efficiency of EAP and MAP is relatively the same while MLE requires more items to estimate someone profile with the ability and standard error that is same. Thus, in the use of CAT with the item selection method based on Efficiency Balanced Information it is recommended to use the Bayesian method (MAP and EAP) in estimating student ability and using the 2PL method in estimating item parameters.

This research is limited to simulation research using WinGen and simulCAT applications. The research focuses on tracking the optimization of EBI as a procedure in selecting items in CAT. To find the optimization of EBI, the researchers combined various parameter estimation methods and methods of estimating students' abilities. It is assumed that all settings in the simulation can run well in the software during the process of generating student ability data and item parameters with WinGen as well as during the CAT implementation simulation using simulCAT.

The results of this study can be used as a basis for implementing EBI as an item selection procedure in CAT. If an agency will develop CAT for various purposes, then it can use EBI as a procedure for selecting items with the parameter estimation method is 2PL and the method of estimating students' ability uses MAP or EAP. However, for more detailed and in-depth results, further research based on original data is needed to demonstrate the original performance of EBI as an item selection procedure in CAT.

REFERENCES

[1] N. Sener and E. Tas, "Developing Achievement Test: A Research for Assessment of 5th Grade Biology Subject," *J. Educ. Learn.*, vol. 6, no. 2, 2017, pp. 254–271.

[2] S. E. R. Kurpius and M. E. Stafford, *Testing and measurement: A user-friendly guide*. Thousand Oaks, California: Sage Publications, Inc., 2006.

[3] H. Retnawati, B. Kartowagiran, J. Arlinwibowo, and E. Sulistyaningsih, "Why are the mathematics national examination items difficult and what is teachers' strategy to overcome it?," *Int. J. Instr.*, vol. 10, no. 3, 2017, pp. 257–276.

[4] N. E. Gronlund, *Constructing Achievement Tests*. Englewood Cliffs, NJ: Prentice-Hall, 1977.

[5] S. A. Sinex, "Investigating Types of Errors," *Spreadsheets Educ.*, vol. 2, no. 1, 2014, pp. 1–10.

[6] J. K. Roberts and R. Herrington, "Demonstration of software programs for estimating multilevel measurement model parameters," *J. Appl. Meas.*, vol. 6, no. 3, 2005, pp. 255–272.

[7] B. A. Bassey, S. V. Ovat, and U. J. Ofem, "Systematic Error in Measurement: Ethical Implications in Decision Making in Learners' Assessment in the Nigerian Educational System," *Prestig. J. Educ.*, vol. 2, no. 1, 2019, pp. 137–146.

[8] H. K. Mohajan, "Two Criteria for Good Measurements in Research: Validity and Reliability," *Ann. Spiru Haret Univ. Econ. Ser.*, vol. 17, no. 4, 2017, pp. 59–82.

[9] R. L. Linn, *Educational Measurement*. NY: American Council on Education and Macmillan, 1989.

[10] J. R. Barrada, J. Olea, V. Ponsoda, and F. J. Abad, "Item selection rules in computerized adaptive testing: Accuracy and security," *Methodology*, vol. 5, no. 1, 2009, pp. 7–17.

[11] F. M. Lord, "Application of Item Response Theory to Practical Testing Problems. Hillsdale, NJ, Lawrence Erlbaum Ass," 1980.

[12] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," *Int. J. Med. Educ.*, vol. 2, 2011, pp. 53–55.

[13] A. J. Nitko and S. M. Brookhart, *Educational assessment of student*. Boston, MA: Pearson Education, Inc., 2011.

[14] M. A. Samsudin, T. Somchut, and M. E. Ismail, "Evaluating computerized adaptive testing efficiency in measuring students' performance in science timss," *J. Pendidik. IPA Indones.*, vol. 8, no. 4, 2019, pp. 547–560.

[15] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory Library*. Newbury Park, California: Sage Publications, 1991.

[16] M. Rezaie and M. Golshan, "Computer Adaptive Test (CAT): Advantages and Limitations," *Int. J. Educ. Investig.*, vol. 2, no. 5, 2015, pp. 128–137.

[17] H. Wainer, E. T. Bradlow, and X. Wang, *Testlet response theory and its applications*. Cambridge: Cambridge University Press, 2007.

[18] A. J. B. Pramono and H. Retnawati, "Implementation of cat in indonesia school: Current challenges and strategies," *Univers. J. Educ. Res.*, vol. 8, no. 11, 2020, pp. 5599–5609.

[19] J. Arlinwibowo, H. Retnawati, B. Kartowagiran, and G. K. Kassymova, "Distance learning policy in Indonesia for facing pandemic COVID-19: School reaction and lesson plans," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 14, 2020, pp. 2828–2838.

[20] M. Marsigit *et al.*, "Constructing Mathematical Concepts through External Representations Utilizing Technology : An Implementation in IRT Course," *TEM J.*, vol. 9, no. 1, 2020, pp. 317–326.

[21] H. Retnawati *et al.*, "Implementing the computer-based national examination in Indonesian schools: The challenges and strategies," *Probl. Educ. 21st Century*, vol. 75, no. 6, 2017, pp. 612–633.

[22] E. Istiyono, W. S. B. Dwandaru, Y. A. Lede, F. Rahayu, and A. Nadapdap, "Developing

IRT-based physics critical thinking skill test: A CAT to answer 21st century challenge," *Int. J. Instr.*, vol. 12, no. 4, 2019, pp. 267–280.

[23] E. Istiyono, W. S. B. Dwandaru, R. Setiawan, and I. Megawati, "Developing of computerized adaptive testing to measure physics higher order thinking skills of senior high school students and its feasibility of use," *Eur. J. Educ. Res.*, vol. 9, no. 1, 2020, pp. 91–101.

[24] T. M. Kantrowitz, C. R. Dawson, and M. S. Fetzer, "Computer Adaptive Testing (CAT): A Faster, Smarter, and More Secure Approach to Pre-Employment Testing," *J. Bus. Psychol.*, vol. 26, no. 2, 2011, pp. 227–232.

[25] D. Magis, D. Yan, and A. A. von Davier, *Computerized adaptive and multistage testing with R*. Cham, Switzerland: Springer, 2017.

[26] H. Wainer, J. D. Neil, R. Flaugher, F. G. Bert, and J. M. Robert, *Computerized adaptive testing: A primer, 2nd edition*. London: Lawrence Erlbaum Associates., 2001.

[27] M. Vrabel, "Computerized versus paper-and-pencil testing methods for a nursing certification examination: A review of the literature," *CIN - Comput. Informatics Nurs.*, vol. 22, no. 2, 2004, pp. 94–98.

[28] M. D. Reckase, "Item Pool Design for Computerized Adaptive Tests," in *The Annual Meeting of the National Council on Measurement in Education*, 2003, no. April, pp. 1–16.

[29] J. van der L. Wim and A. W. G. Gees, *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer Academic, 2000.

[30] T. J. H. M. Eggen and G. J. J. M. Straetmans, "Computerized adaptive testing for classifying examinees into three categories," *Educ. Psychol. Meas.*, vol. 60, no. 5, 2000, pp. 713–734.

[31] T. Wang and W. P. Vispoel, "Properties of ability estimation methods in computerized adaptive testing," *J. Educ. Meas.*, vol. 35, no. 2, 1998, pp. 109–135.

[32] K. T. Han, "An Efficiency Balanced Information Criterion for Item Selection in Computerized Adaptive Testing," *J. Educ. Meas.*, vol. 49, no. 3, 2012, pp. 225–246.

[33] N. A. Thompson and D. J. Weiss, "A framework for the development of computerized adaptive tests," *Pract. Assessment, Res. Eval.*, vol. 16, no. 1, pp. 1–9, 2011.

[34] A. Birnbaum, "Some latent trait models and their use in inferring an examinee's ability," in *Statistical theories of mental test scores*, F. M. Lord and M. R. Novick, Eds. MA: Addison-Wesley, 1968, pp. 397–472.

[35] W. J. Van Der Linden and R. K. Hambleton, *Handbook of modern item response theory*. NY: Springer, 1996.

[36] W. J. van der Linden and P. J. Pashley, "Statistics for Social and Behavioral Sciences," in *Element of adaptive testing*, W. J. van der Linden and A. W. G. Cees, Eds. 233 Spring Street, NY: Springer, 2010, pp. 3–30.

[37] C. DeMars, *Item response theory*. .Madison Avenue, New York: Oxford University Press, Inc, 2010.

[38] H. Retnawati, "Perbandingan Estimasi Kemampuan Laten Antara Metode Maksimum Likelihood Dan Metode Bayes [Latent Capability Between Maximum Likelihood Method And Bayes Method]," *J. Penelit. dan Eval. Pendidik.*, vol. 19, no. 2, 2015, pp. 145–155.

[39] M. D. Reckase, *Multidimensional item response theory*. Spring Street, New York: Springer, 2009.

[40] F. B. Baker, *The basics of item response theory*, (2nd ed.). College Park, MD: ERIC, 2001.

[41] J. Choi, S. Kim, J. Chen, and S. Dannels, "A comparison of maximum likelihood and bayesian estimation for polychoric correlation using Monte Carlo simulation," *J. Educ. Behav. Stat.*, vol. 36, no. 4, 2011, pp. 523–549.

[42] S. Y. Phoong and M. T. Ismail, "A comparison between Bayesian and maximum likelihood estimations in estimating finite mixture model for financial data," *Sains Malaysiana*, vol. 44, no. 7, 2015, pp. 1033–1039.

[43] B. Pandey, N. Dwividi, and B. Pulastya, "Comparison between bayesian and maximum likelihood estimation of the scale parameter in Weibull distribution with known shape under linex loss function," *J. Sci. Res.*, vol. 55, no. 1, 2011, pp. 163–172.

[44] A. F. Borgatto, C. Azevedo, A. Pinheiro, and D. Andrade, "Comparison of ability estimation methods using IRT for tests with different degrees of difficulty," *Commun. Stat. Simul. Comput.*, vol. 44, no. 2, 2015, pp. 474–488.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[45] C. G. Parshall, J. A. Spray, J. C. Kalohn, and T. Devey, *Practical consideration in competur-based testing.* NY: Springer, 2012.

[46] B. M. Triyono, N. Mohib, G. K. Kassymova, G. N. I. P. Pratama, D. Adinda, and M. R. Arpentieva, "The profile improvement of vocational school teachers' competencies," Vyss. Obraz. v Ross., vol. 29, no. 2, 2020, pp. 151–158.