REiD (Research and Evaluation in Education), 6(1), 2020, 32-40
Available online at: http://journal.uny.ac.id/index.php/reid

Alternative item selection strategies for improving test security in computerized adaptive testing of the algorithm

[*]Iwan Suhardi
Faculty of Engineering, Universitas Negeri Makassar
Jl. Daeng Tata Raya, Parang Tambung, Mannuruki, Tamalate, Kota Makassar, Sulawesi Selatan 90224, Indonesia
[*]Corresponding Author. E-mail: iwan.suhardi@unm.ac.id

Submitted: 5 March 2020 | Revised: 21 April 2020 | Accepted: 29 April 2020

Abstract

One of the ability estimation methods that is widely applied to the Computerized Adaptive Testing (CAT) algorithm is the maximum likelihood estimation (MLE). However, the maximum likelihood method has the disadvantage of being unable to find a solution to the ability estimation of test-takers when the test takers' scores do not have a pattern. If there are test takers who get either score of 0 or perfect score, then the abilities of test-takers are usually estimated using the step-size model. However, the step-size model often results in item exposure where certain items will appear more often than other items. This surely threatens the security of the test because items that often appear will be easier to recognize. This study tries to provide an alternative strategy by modifying the step-size model and randomizing the calculation results of the information function obtained. Based on the results of the study, it is found that alternative strategies for item selection can make more varied items appear to improve the security of tests on the CAT.

Keywords: *item selection strategy, item exposure, step-size, adaptive testing*

How to cite: Suhardi, I. (2020). Alternative item selection strategies for improving test security in computerized adaptive testing of the algorithm. *REiD (Research and Evaluation in Education), 6*(1), 32-40. doi:https://doi.org/10.21831/reid.v6i1.30508.

Introduction

The development of item response theory (IRT) and computer technology that is faster and in a large capacity allows the development of computerized adaptive testing (CAT) (Haryanto, 2013, pp. 49–50). It is called "computerized" testing because the testing process no longer uses paper and pencil, but rather uses a computer device. It is called "adaptive" testing because the items that appear are chosen in such a way and adjusted to the ability of the test takers independently. CAT is a test conducted for test-takers where the items are determined based on the answers of the test takers (Winarno, 2013, p. 577). The efficiency of CAT compared to conventional testing models has been supported by several studies. The results of research by Eignor concluded that at the same level of measurement precision, adaptive tests only required a test length that was less than half of the computer-based test (CBT) device (Eignor, Stocking, Way, & Steffen, 1993; Grist, 1989, p. 2; Rudner, 1998, p. 2). McBride and Martin concluded that to achieve the same level of reliability, conventional testing required 2.57 times more items than adaptive testing (McBride & Martin, 1983).

The method widely used to estimate the ability of test-takers is the maximum likelihood estimation (MLE). The application of

the maximum likelihood method has the disadvantage of being unable to find a solution when there are test takers who get extreme scores where all answers are always incorrect or always correct. To overcome this problem, the step-size method is generally employed. However, the application of the MLE and step-size model often leads to item exposure, which is the frequent appearance of certain items given to test takers. Although CAT is more efficient and reliable, the security of this testing is not guaranteed because certain items appear repeatedly. The items are easily recognized because they appear frequently, especially at the beginning of the item sequence. Therefore, modifications are needed to the conventional CAT algorithm to minimize the appearance of these easily noticeable items. The procedures that are commonly used in developing conventional CAT algorithms are elaborated as follows (Thissen, 1990).

Starting CAT

CAT generally starts with the selection of items with the difficulty level of moderate (Mills, 1999, p. 123; Santoso, 2010, p. 70; Vispoel, 1999). A test taker who answers incorrectly will then be given items with the difficulty level of easy. Conversely, if test taker answers correctly, they will be given items with the difficulty level of hard.

Estimating the Ability of the Test-Takers

The method commonly used to estimate the ability of test-takers is MLE (Baker, 1992; Birnbaum, 1968). The estimation of the ability of test-takers using the maximum likelihood method is calculated using the Newton-Raphson iterative procedure (Hambleton & Swaminathan, 1985, p. 83). The Newton-Raphson iterative procedure is performed first by subtracting the ratio of the first derivative to the second derivative from the initial $\hat{\theta}$ value so that it results in new $\hat{\theta}$. This procedure is repeated by using the new $\hat{\theta}$ and calculating the value of the new derivative ratio. The estimated value of θ at (m + 1) iteration can be expressed using the iterative relation as presented in Formula (1). Meanwhile, the error value is a correction factor that is formulated as seen in Formula (2), where *u* equals 1 if student's answer is correct and *u* equals 0 if student's answer is incorrect. Besides, *P* is probability of participants answering the items correctly, which is obtained by Formula (3).

$$\theta_{m+1} = \theta_m + error \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

$$error = \frac{\sum 1.7\, a\, (u-P)(P-c)/(P(1-c))}{\sum [-1.7^2\, a^2\, ((1-P)/P)]\,[(P-c)/(1-c)]^2} \quad \ldots \quad (2)$$

$$P = c + \frac{(1-c)}{\left(1+exp\left(-1.7\, a\, (\theta_{duga\,0}-b)\right)\right)} \quad \ldots. \quad (3)$$

The iteration process is stopped when the error value $< \varepsilon$, with ε as limiting number whose value is very small. In this study, the ε value of 0.0001 was used.

One problem with the application of the MLE method in adaptive testing is the inability of the MLE method to find solutions when there are test takers who get an extreme score, which is either a score of 0 or a perfect score. To overcome the problem of the inability of the MLE method to estimate the ability of test-takers when their responses did not have a pattern, the step size method can be used (Dodd, 1990). Based on the step size method, the test taker's ability level is upgraded or degraded by a certain constant as long as the test taker's responses do not have a pattern, for example, by using a step size of 0.5.

Selection of the Next Item

After the test taker's ability is successfully estimated, the CAT algorithm will then select the next item. Lord recommended the use of the maximum item information procedure to select the next item (Lord, 1977). This method guarantees a highly accurate estimation of the ability of test-takers (Eignor et al., 1993). Items that have the greatest information function value on the ability of certain test takers are selected to be presented to them. The item information function is calculated at each ability level with the equation in Formula (4) (Hambleton, Swaminathan, & Rogers, 1991, p. 107).

$$I(\theta) = \frac{2.89\, a_i^2\,(1-c_i)}{\left[\left(c_i + exp(1{,}7\, a_i(\theta-b_i))\right)\right]\left[1+exp(-1{,}7\, a_i(\theta-b_i))\right]^2} \quad \ldots. \quad (4)$$

Formula (4) shows that the information value only depends on the characteristic value of item parameters (for example the values of b, a, and c for the 3PL model) and the level of ability (θ). Thus, for each ability level (θ), the information function contribution for each item in the question bank can be calculated.

The test information function is the sum of the information functions of the test item and is written as in Formula (5). Meanwhile, the test information function illustrates the accuracy of the test set in estimating different levels of ability. The greater the information at the given ability level, the more accurate the ability is estimated from the test kit. The standard error of measurement (SEM) is expressed by the equation in Formula (6) (Hambleton & Swaminathan, 1985, p. 95).

$$TIF = \sum_{i=1}^{n} I_i \quad \text{..... (5)}$$

$$SEM = {}^{1}/_{\sqrt{TIF}} \quad \text{..... (6)}$$

Termination of CAT

CAT termination uses criteria of equal measurement precision and a fixed number of items. Equal measurement precision criteria aim to produce test scores with the same measurement error level for each test taker. The standard error of measurement is limited to 0.30, which is equivalent to reliability of 91% on conventional tests (Thissen, 1990). By using the criteria, the number of items the test takers must work on can vary (where the number of items is not the same). However, to avoid the test process that may not be converging, the criterion of a fixed number of items is also used in the CAT termination rules by limiting the maximum items that appear, for example, as many as 20 items.

Giving Score to the Ability of the Test-Takers

The score of the ability estimation of the test-taker derives from the conversion of the value θ that is obtained by Formula (7).

$$Score = 50 + \left(\frac{50}{3}\theta\right) \quad \text{..... (7)}$$

In this study, the CATs assessment results, which were the conventional CAT model (by taking the information value of items or the largest I (θ)) and the alternative CAT model (by taking some of the largest I (θ) values, then taken randomly to determine the value of I (θ) that would be used), were compared. After that, the alternative CAT model was treated using the step-size method with an additional variable of response time when the test takers' responses did not have a pattern.

The assumption underlying the response time variable is those test-takers who have a high level of ability will be able to answer the items correctly in a shorter time than those who have a lower level of ability. Lidia Martinez compared groups of test-takers who took a test using CBT and found that the groups that spent the shortest average time responding to the initial test item obtained a higher average score (Martinez, 2009). Phil Higgins' research results showed that in CBT, if the item difficulty index was higher, then test-takers would need more time to answer and review the items (Higgins, 2009). This showed that the test taker's response time in working on the items correctly correlated with the estimation of the test taker's ability level.

Method

This study used a Research and Development (R&D) approach. The study began with the development of a question bank to obtain 265 items based on the 1-parameter logistic item response theory (1PL IRT) model. Characteristics of items in the form of parameters of the difficulty level of 265 items were obtained from the validation of processed results using the BILOG-MG software, obtained from the response test using CBT. The total number of items before validation was originally 290 items. A summary of the question bank validation statistics developed and used in this study is presented in Table 1.

Table 1. Summary of Item Statistics on Question Bank

General Information	Based on 1PL IRT Number of items = 265 items
Criteria of Item Difficulty Index (*b*)	Hard category = 40 items Moderate category = 128 items Easy category = 97 items

In the 1PL IRT model, the probability of a person with a certain ability (θ) answering the items correctly depends only on the difficulty level of the items (b). In this study, the estimation methods of the ability of test-takers are the MLE and step-size methods.

Next, two adaptive test designs developed were the conventional and the alternative CAT model. In this study, the development of CAT software referred to the incremental model (Pressman, 2001, pp. 35–36).

In the conventional CAT model, the first item selection method employs a difficulty level of moderate, starting with a range of b values from -0.5 to 0.5 chosen randomly. The ability level estimation is calculated using the MLE method. However, when the test-takers' responses have not had a pattern, their ability is estimated using the step size method with a value of 0.5. The next item that is selected is the item that has the greatest information function value on a particular ability.

The alternative CAT model has the same principles as the conventional CAT model. The difference is in the selection of the second and subsequent items, which uses the principle of randomizing the value of the information function in the 5-4-3-2-1 pattern. The pattern rule of 5-4-3-2-1 used was that the second item was selected from one item randomly from the five items that had the largest information function, the third item was selected from one item randomly from the four items that had the largest information function, the fourth item was selected from one item randomly from three items that had the largest information function, the fourth item was selected from one item randomly from three items that had the largest information function, and the fifth item was selected from one item randomly from two items that had the largest information function. Meanwhile, for the sixth and subsequent items, the item selection criteria revert to the maximum information function criteria or revert to the conventional CAT model.

To estimate the ability of test-takers on the alternative CAT model when their responses have not had a pattern, a step-size method is used with the addition of the response time variable. The test-takers' estimated initial ability level is selected at the ability level of $\theta 0$. Moreover, the step-size interval changes constantly by $\mathbf{k}$ (where in this study, the value of $\mathbf{k}$=0.5). If the test taker responds by answering incorrectly, the test-taker's estimated ability level becomes $\boldsymbol{\theta 0 - k}$ or equal to 0-0.5 = -0.5. Meanwhile, if the test taker answers correctly, the estimated ability level becomes $\boldsymbol{\theta 0 + x\,k}$ or $\mathbf{0.5}\,.\,\boldsymbol{x}$, where $\boldsymbol{x}$ is a positive constant multiplier and the value depends on the category of students' response time when their answer is correct.

Table 2 shows a simulation procedure to estimate the test taker's ability level with a step-size interval added to the response time factor. Test takers were given 300 seconds to respond to each item. If for more than 300 seconds there is no response from test taker, the response is declared incorrect and easier-level items will be displayed. In this study, the criterion for test termination is that the test is terminated if the SEM value has reached 0.30. An SEM value of 0.30 is equivalent to the reliability of 0.91 in conventional tests such as paper and pencil tests (Thissen, 1990).

Table 2. Estimation of Ability of Test-Taker in the Response-Time-Based Step-Size Method

Annotation: $\boldsymbol{\theta_0}$ = Initial ability = 0 **k** = step size = 0.5 **x** = constant multiplier $\boldsymbol{\theta_{ke\text{-}i} = \theta_{i\text{-}1} + xk}$ (for correct response) $\boldsymbol{\theta_{ke\text{-}i} = \theta_{i\text{-}1} - k}$ (for incorrect response)	**Responding with Correct Answer in Consecutive Times**			**Responding with Incorrect Answer in Consecutive Times**		
	Item 1	**Item 2**	**Item 3**	**Item 1**	**Item 2**	**Item 3**
	$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	$\boldsymbol{\theta_3}$	$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	$\boldsymbol{\theta_3}$
Very fast: x = 1.4 (≤ 30 seconds)	0.7	1.4	2.1	-0.5	-1.0	-1.5
Fast : x = 1.3 (31 to 60 seconds)	0.65	1.3	1.95	-0.5	-1.0	-1.5
Moderate: x = 1.2 (61 to 90 seconds)	0.6	1.2	1.8	-0.5	-1.0	-1.5
Slow : x = 1.1 (91 to 120 seconds)	0.55	1.1	1.65	-0.5	-1.0	-1.5
Very slow : x = 1 (≥ 121 seconds)	0.5	1.0	1.5	-0.5	-1.0	-1.5

Table 3. Testing Results of Conventional CAT Model when Responses of Answers Have Not Had Pattern Yet

Item 1	Item 1 was taken randomly with the difficulty level of moderate (-0.5 ≤ b ≤ 0.5)			
Item	**Test Takers' Responses are Always Correct**		**Test Takers' Responses are Always Incorrect**	
	Value of θ	**List Number of Item**	**Value of θ**	**List Number of Item**
Item 2	- 0.5	209	0.5	275
Item 3	- 1	164	1	081
Item 4	- 1.5	113	1.5	002
Item 5	- 2	044	2	091
Item 6	- 2.5	237	2.5	115

Findings and Discussion

Before the answers have a pattern, the conventional CAT model will use the step-size method with an interval of 0.5. This means that if the test taker always responds with the correct answer, then the second and subsequent items that will appear are items that have the largest information function value at the ability level (θ) of 0.5, 1, 1.5, 2, 2.5, and 3 respectively. Meanwhile, for test-takers who always respond with the incorrect answer, the second and subsequent items that will appear are items that have the largest information function value at θ of -0.5, -1, -1.5, -2, -2.5, and -3 respectively. The results that were obtained in the conventional CAT model are summarized in Table 3.

From the results of the study, it was found that items with list numbers of 209, 164, 113, 044, 237, 275, 081, 002, 091, and 115 were items that appeared more often than other items. The items that often appear will make the security of the test in the conventional CAT model degrade because they may be items that have been recognized by the test takers.

From the results of conventional CAT model testing, it was found that the number of items with difficulty index of moderate, which was indicated by the difficulty index value (b) ranging from -0.5 to +0.5, was 128 items. This meant that the probability of the first item having a chance to appear was 128 items chosen randomly. This was indeed in accordance with the criteria applied to the conventional CAT model design algorithm, that the initially selected items were items with difficulty index of moderate (-0.5 to +0.5).

After the first item displayed and was responded by the test taker, the second item was presented by using the step-size method. This meant that if students responded to the item with the correct answer, then the second item displayed was the item with maximum information for θ = 0.5. However, if students responded to items with incorrect answers, then the second item that was displayed was an item with maximum information for θ = -0.5. Thus, it was certain that in the conventional CAT, the second item only consisted of the possibility of 1 of 2 items only. In this study, the second item presented was question item number 275 (if the answer was correct) and question item number 209 (if the answer was incorrect). The frequent appearance of item number 275 and item number 209 made the security of CAT threatened due to the familiarity with the question.

Another case that also often arises is that there has not been a pattern in students' answers so that the step-size method is used. For example, if students answered questions correctly, the items that would appear were questions that had a maximum information value for θ = 0.5, 1.0, 1.5, 2.0, and 2.5, which were the second item whose item number was 275, the third item whose item number was 081, the fourth item whose item number was 002, the fifth item whose item number was 091, and the sixth item whose item number was 115.

However, if students always answered the question incorrectly, then the item that appeared was questions that had a maximum information value for θ = -0.5, -1.0, -1.5, -2.0, and -2.5, i.e., the second item with item number 209, third item with item number 164, fourth item with item number 113, fifth item with item number 044, and sixth item with item number 237. In the conventional CAT model, if the responses of the test takers

have the same pattern, then the items that appear will also be the same. This is what makes the security level of the conventional CAT model suboptimal.

If students' responses already had patterns (where the responses already consisted of correct and incorrect answers), then the items that appeared next had been quite varied because the first item that appeared already had a relatively large variety of items (128 items). However, by using the maximum information function value model to search for items that corresponded to the estimated level of test-takers' abilities, it was very possible that many items could not be presented because they never obtained the maximum function value for each level of ability.

The alternative solution proposed was to use the step-size method based on the student's response time in answering correctly. Student responses were grouped into groups based on the time spent by students in answering the questions correctly. In the step-size method based on response time, the step-size value formula was given an additional constant multiplier based on the response time. The faster the students answered correctly, the greater the constant multiplier became.

An additional solution proposed was to randomize the maximum information function value. If the conventional CAT model determined the items that appeared based on the value of the (single) maximum information function, then the alternative CAT model determined the items that appeared by randomizing the maximum information function values based on groups of 5–4–3–1–1. For example, one of the results of testing the alternative CAT model is presented in Table 4.

From Table 4, the calculation procedure for the alternative CAT model can be observed. From the table, it can be seen that the items that appear in the alternative CAT model are more varied compared to those in the conventional CAT model. The algorithmic procedure in the alternative CAT model can be explained as follows.

The First Item that Appeared was Item Number 239 with b = -0.416

The first item appeared in accordance with the criteria that items were taken randomly with a difficulty index of moderate whose b value ranged from -0.5 to 0.5. Item number 239 fulfilled the criteria. Because students' answers did not have a pattern, the method of estimating the ability level was the step-size of 0.5. Students' answers were declared correct (value 1). The time that was spent to work on the first item was 34 seconds, so it was included in the fast category (between 31 and 60 seconds) with a multiplier factor = 1.3. Thus, the value of θ was 0.5 x 1.3 = 0.64.

Table 4. Results of Alternative CAT Model Testing

No.	Item	*b*	Response	Time (second)	θ	IIF	TIF	SEM
1	239	-0.416	1	34	0.65	0.7224	0.7224	1.18
2	182	0.662	1	40	1.3	0.7223	1.4447	0.83
3	192	1.32	0	8	1.1809	0.7225	2.1672	0.68
4	042	1.181	0	49	0.8579	0.7225	2.8897	0.59
5	132	0.861	1	20	1.3204	0.7225	3.6122	0.53
6	192	1.32	0	26	1.1161	0.7225	4.3347	0.48
7	152	1.119	1	10	1.5224	0.7225	5.0572	0.44
8	002	1.524	0	14	1.3846	0.7224	5.7796	0.42
9	161	1.396	0	7	1.2399	0.7224	6.502	0.39
10	013	1.251	1	9	1.5831	0.7225	7.2245	0.37
11	127	1.579	1	15	1.9486	0.7217	7.9462	0.35
12	060	1.987	0	12	1.8848	0.7223	8.6685	0.34
13	062	1.867	0	17	1.8118	0.7222	9.3907	0.33
14	163	1.787	0	19	1.7339	0.7214	10.1121	0.31
15	124	1.687	1	14	2.0656	0.7214	10.8335	0.3

The Second Item that Appeared was Item Number 182 with $b = 0.662$

The second item appeared because it had the five largest information function values at the value of θ = 0.65, according to the use of randomization with the principle of 5–4–3–2–1. From the five alternative values of the largest information function (see Table 5), the item with number 182 was selected randomly. The second item was answered correctly (then the response value was 1). Because students' answers did not have a pattern, the method of determining the estimated ability level was the step-size of 0.5. The item was done in 40 seconds and included in the fast category (between 31 to 60 seconds) with a multiplier factor of 1.3. Thus, the value of θ = 0.65 + (0.5 x 1.3) = 1.3.

Table 5. The Five Alternative Values of the Largest Information Function

Rank	Information Function	Item	*b*
1	0.722495	153	0.647
2	0.722492	274	0.654
3	0.722474	202	0.643
4	0.722425	182	0.662
5	0.721861	003	0.685

The Third Item that Appeared was Item Number 192 with $b = 1.32$

The third item appeared because it had the four largest information function values at the value θ = 1.3 according to the use of randomization with the principle of 5–4–3 –2–1. Of the four alternative values for the largest information function (see Table 6), the item with number 192 was randomly selected. The third item was responded with an incorrect answer (so the response value was 0). Because students' answers did not have a pattern, the method for estimating the level of ability was MLE. The value of θ obtained was = 1.1809.

Table 6. The Four Alternative Values for the Largest Information Function

Rank	Information Function	Item	*b*
1	0.722349	053	1.317
2	0.722291	192	1.32
3	0.72227	179	1.321
4	0.722091	145	1.272

The Fourth Item that Appeared was Item Number 042 with $b = 1.181$

The fourth item appeared because it had the three largest information function values at the value θ = 1.1809 according to the use of randomization with the principle of 5–4–3 –2–1. Of the three alternative values for the largest information function (see Table 7), item with number 042 was randomly selected. The fourth item was responded with an incorrect answer (so the response value was 0). Because students' answers did not have a pattern, the method for estimating the level of ability was MLE. The value of θ obtained was = 0.8579.

Table 7. The Three Alternative Values for the Largest Information Function

Rank	Information Function	Item	*b*
1	0.7225	042	1.181
2	0.7225	057	1.181
3	0.722449	021	1.171

The Fifth Item that Appeared was Item Number 132 with $b = 0.861$

The fifth item appeared because it had the two largest information function values at the value θ = 0.8579 according to the use of randomization with the principle of 5–4–3 –2–1. Of the two alternative values for the largest information function (see Table 8), the item with number 132 was randomly selected. The fifth item was responded with the correct answer (so the response value was 1). Because students' answers did not have a pattern, the method for estimating the level of ability was MLE. The value of θ obtained was = 1.3204.

Table 8. The Two Alternative Values for the Largest Information Function

Rank	Information Function	Item	*b*
1	0.722495	132	0.861
2	0.722474	242	0.865

The Sixth Item that Appeared was Item Number 192 with $b = 1.32$

This sixth item appeared because it had one largest information function value at the value θ = 1.32 according to the use of ran-

domization with the principle of 5–4–3 –2–1. Of the one alternative value for the largest information function (see Table 9), the item with number 192 was randomly selected. The sixth item was responded with an incorrect answer (so the response value was 0). Because students' answers were patterned, the method for estimating the level of ability was MLE. The value of θ obtained was = 1.1161.

Table 9. The One Largest Information Function Value

Rank	Information Function	Item	*b*
1	0.7225	192	1.32

The subsequent items (i.e. the seventh to fifteenth items) used the same method to determine the item that had the largest information function at its value of θ. The fifteenth item became the last item because the criterion for termination rule had been met (SEM = 0.3). It was converted to a numerical value of 85.

This alternative CAT model has been proven to be able to overcome a fundamental shortcoming in the conventional CAT model, which was the frequent appearance of certain items. From Table 3, it can be seen that in the conventional CAT model, several similar items would appear, especially in the initial patterns of CAT execution. Meanwhile, in Table 4, there were many variations on the possible items that appeared on the alternative CAT model, even though the patterns of students' answers were the same. The many variations of items that appear in the alternative CAT model can reduce the level of item exposure on CAT so that it will make the CAT more secure. The item variations that appeared in the alternative CAT model actually had item difficulty index that was not much different from those that appeared in the conventional CAT model, so it did not increase the test length or reduce the efficiency of the estimation of the ability of the test takers.

Conclusion

From the results of this study, it can be concluded that the alternative CAT model was able to decrease the level of item exposure on the CAT, thereby increasing the security of the CAT without increasing the test length or reducing the efficiency of the CAT. The strategy adopted by the alternative CAT model was to select items using the step-size method based on response time and randomization of the maximum information function value with the criteria of 5–4–3–1–1 by applying the maximum likelihood estimation (MLE) to estimate the ability level of the test takers. The strategy has been proven to be able to present items with more variations, but still with item difficulty index which was not much different in the response patterns of the same test takers.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental rest scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement*, *14*(4), 355–366. https://doi.org/10.1177/014662169001400403

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation*. https://doi.org/10.1002/j.2333-8504.1993.tb01567.x

Grist, S. (1989). Computerized adaptive tests. In *ERIC Digest No. 107*. Retrieved from https://files.eric.ed.gov/fulltext/ED315425.pdf

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item*

response theory. Newbury Park, CA: Sage Publications.

Haryanto, H. (2013). Pengembangan computerized adaptive testing (CAT) dengan algoritma logika Fuzzy. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *15*(1), 47–70. https://doi.org/10.21831/pep.v15i1.1087

Higgins, P. (2009). *Candidate measured ability and use of time*. Retrieved from https://www.rasch.org/mra/mra-10-09.htm

Lord, Frederic M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, *1*(1), 95–100. https://doi.org/10.1177/014662167700100115

Martinez, L. (2009). *Time usage and candidate performance*. Retrieved from http://www.rasch.org/mra/mra-06-09.htm

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 224–236). New York, NY: Academic Press.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examinations General Test. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117–135). Mahwah, NJ: Lawrence Erlbaum Associates.

Pressman, R. S. (2001). *Software engineering: A practitioner's approach* (5th ed.). New York, NY: McGraw-Hill Higher Education.

Rudner, L. M. (1998). *An on-line, interactive, computer adaptive testing tutorial*. Retrieved from http://edres.org/scripts/cat

Santoso, A. (2010). Pengembangan computerized adaptive testing untuk mengukur hasil belajar mahasiswa Universitas Terbuka. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *14*(1), 62–83. https://doi.org/10.21831/pep.v14i1.1976

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 161–186). Hillsdale, NJ: Erlbaum.

Vispoel, W. P. (1999). Creating computerized adaptive tests of music aptitude: Problems, solutions, and future directions. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 151–176). Mahwah, NJ: Lawrence Erlbaum Associates.

Winarno, W. (2013). Pengembangan computerized adaptive testing (CAT) menggunakan metode pohon segitiga keputusan. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *16*(2), 574–592. https://doi.org/10.21831/pep.v16i2.1132