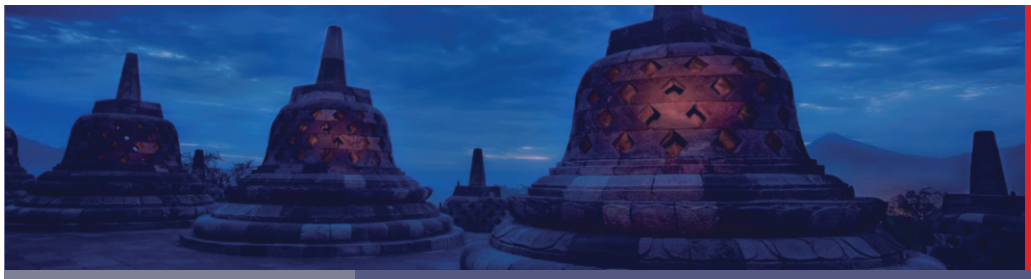

ISSN: 2407-1501

PROCEEDING

INTERNATIONAL CONFERENCE ON EDUCATIONAL RESEARCH AND EVALUATION (ICERE)

"Assessment for Improving Students' Performance"

May 29 – 31 2016
Rectorate Hall and Graduate School
Yogyakarta State University
Indonesia

Organized by:
Study Program of Educational Research and Evaluation
Graduate School, Yogyakarta State University
in Cooperation with Indonesian Educational Evaluation Association (HEPI),
and Center for Educational Assessment (PUSPENDIK) Ministry of Education and Culture

MODIFICATION OF RANDOMIZED ITEMS SELECTION AND STEP-SIZE BASED ON TIME RESPONSE MODEL TO REDUCE ITEM EXPOSURE LEVEL OF CONVENTIONAL COMPUTERIZED ADAPTIVE TESTING

Iwan Suhardi
Department of Electrical Engineering, Faculty of Engineering
State University of Makassar, Makassar, Indonesia
email: iwansuhardi@yahoo.com

Abstract - To estimate the ability of students, computerized adaptive test (CAT) has been shown to be more efficient than conventional tests using paper and pencil test (P&P Test) and computerized-based testing (CBT). However, the conventional CAT applying the Maximum Likeli-hood Estimation (MLE) and the step-size to estimate the ability of the test taker and the application of item information function (IIF) for the selection of items displayed will cause item exposure problems or frequent emergence of some items that given to test participants, so that the item was easy to spot, especially at the beginning of the emergence in the order items. This paper attempts to determine the effect modification by randomization in the CAT algorithm and step-size based on the response time to estimate the ability of the test taker. Items bank for the research using item response theory models one parameter logistic (1 PL). Development model is the method of randomization using 5-4-3-2-1 models based on MLE and grouping the response time for a constant step-sizenya. Based on study results, the CAT algorithm modification resulted in the appearance items are more varied, thereby reducing exposure item problem without reducing the efficiency of CAT.

Keywords*: computerized adaptive test, item exposure, randomization, step-size*

I. INTRODUCTION

Generally, the test was built to estimate the ability of participants test. Giving the test too easy for the person taking the test is a waste of time and otherwise, the questions that are too difficult, also produces test scores that are not informative. For customizing a test to bring the level of ability of each individual participant test, a solution should be sought. In the development of modern measurement theory, item response theory (IRT) as well as the advances in computer technology makes it possible to develop a computerized adaptive test, or more popularly known as the Computerized Adaptive Testing (CAT) [1] [2]. Known as “computerized”, because the implementation of computerized testing really is no longer using paper and pencil. “Adaptive”, because items have been selected based on the results of the self-regulatory analysis and adapted to the needs or abilities of the examinees, works automatically through a computer software. According to Weiner, CAT is a test held for participants where the items are determined based on the response from the participants’ answer [2].

Comparison between traditional testing and adaptive testing in Table 1 below [3]:

Table 1. Comparison Between Traditional And Adaptive Testing

FACTOR	TRADITIONAL TESTING	ADAPTIVE TESTING
Composition Test	Each participant received a similar test	Each participant received a different test
Difficulties Test	Intended for the average participant	Intended for individual participants
Length Test	Identical for all participants of the test. In general, a long test	The length of the test is not the same for all participants. In general, a short test
Time Test	Certain time	Anytime
Organization Test	It takes a long time	It takes a short time
Results Instantly	Generally, it requires a long time to see results	The results appear instantly

Efficiency of CAT is supported by several studies. McBride and Martin concluded that to achieve the same level of reliability, the conventional tests still require as much as 2.57 times the number of items in adaptive test [4]. Similar research results by Eignor also concluded that the adaptive test only requires a long test less than half of the length of the paper and pencil test at the same level of precision measurements [5].

However, application of the maximum likelihood method and step-size at CAT for estimating the ability of test takers have a number of items given to participants of the test appear more often compared other items. This occurs especially at the beginning of the emergence of items given to participants of the test. Therefore, it is necessary to modify the CAT algorithm to reduce the problem of often appear items and are easily recognizable. This is known as item exposure. Although the design of adaptive test is more efficient and reliable, it is not guaranteed as safety testing because of frequent emergence of certain items.

II. THEORY OF CAT

The practical steps commonly used to develop conventional CAT algorithm are as follows [6]:

A. CAT Starting Point

If no preliminary information about the ability of partisipants, the CAT can begin by selecting the items beginning with a medium level of difficulty [7] [8].

B. Continuing Process

After obtaining the response of the participants' answers, CAT system gives a response assessment with a correct or incorrect answer. There are two steps to continue the process of CAT, which estimate the ability of the participants and how to choose the next items.

1. Method of Estimating Capabilities

Having answered the first item given, the ability of the test taker is estimated based on the parameters of items, the estimated value of the initial capabilities, and answers of the items whether true or false. The general method used to estimate the ability of the test taker is Maximum Likeli-hood Estimation (MLE) [9] [10]. One problem with the application of MLE method on adaptive testing is the inability of the likelihood function to find a solution when there is a maximum of examinees who earn a score of 0 (answered incorrectly on all items) or a perfect score (answered correctly on all items). To overcome the problem of the inability of MLE method in estimating the level of response capability when the participants have not figured test participants can use the method step size [11] [12]. Based on the method step size, ability level test participants increased or decreased by a certain number during the test have not been patterned response.

2. Selection of Next Items

Once the ability of participants is estimated, the computer select the next items. Lord suggests using items maximum information procedure to select the next items to be given to the participants of the test [13]. Based on this procedure, item that has a greatest value information function on the ability of certain participants have to be given to the test taker. This ensures that the value of the function test information for each person taking the test is maximum, meaning that the standard error of measurement (SEM) minimum because no other test information function is the inverse of the variance of the measurement error. In other words, this method

guarantees yield prediction skill level of participants with high accuracy [5]. Using the information function, the accuracy of measurement in estimating the ability of test takers can be calculated at every level of ability. Function Birnbaun information item to be stated by the following equation [14]:

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

The above equation shows that the information is only dependent on item parameter (eg a, b, and c for the model 3P) and the level of ability (θ). Test information function is the number of item information function test developers [15]. Information test function device is mathematically written as follows:

$$TIF = \sum_{i=1}^{n} I_i \quad (2)$$

As an item information function, the test information function illustrate how accurate the estimate test for different ability levels. The greater level of information on given ability, the more accurate the estimated ability of the test device. Standard error of measurement (SEM) is expressed by the following equation [15]:

$$SEM = {}^{1}/_{\sqrt{TIF}} \quad (3)$$

C. Stopping Rule

Two main methods are used to stopping CAT, equal measurement precision and fixed number item. Both of these methods produce different measurement error variance. The purpose of the equal measurement precision method is generating test scores with the same error rate measurements for each test taker's ability. Standard error of measurement equivalent set a limit on 0.30 with a reliability of 91% on conventional tests [16]. But in practice it is also use criteria fixed number of items, the dismissal rules CAT, eg using criteria fixed starting rule as much as 20 items to avoid the process of tests that may not converge.

In this study, two draft adaptive test developed are a conventional CAT (not randomized) and a randomized. The design of a randomized CAT is principally the same as the design of conventional adaptive tests. The difference lies in the selection of items for second item and subsequent use of randomization principle 5 – 4 – 3 – 2 – 1. To estimate the ability of the test taker, when the response of participants has not been patterned, used the method step-size based on the response time.

Participants test that have a high skill level is assumed to be able to answer the item correctly in a faster time than the learners who have less ability levels. Use of the speed test participants' responses factor to the additional assessment information is also recommended by Dunkel [17]. Van der Linden said that if the speed of response and accuracy-related or if both are important in the context of the test, the speed of response can be included in the assessment rubric [18]. Lidia Martinez' research on CBT found that the group spend an average time to respond to the fastest initial test items have an average score higher. However high or low the average score is statistically not influenced by the length of time the person taking the test to review the previous item [19]. This indicates that the speed in responding to the items correctly influenced by the ability of the test taker.

The results of the research are almost the same also delivered by Phil Higgins. Group of test participants with high scores able to complete the items properly with the average time that is shorter than the test group of participants with moderate and low scores [20]. In another study on the CBT, Higgins also found that higher levels of item difficulty, the person taking the test will need more time to answer and review the item [21]. This shows that the response time test takers work item correctly correlated with the level of ability of the test taker.

Chang's research concluded that there was no statistical difference with regard to gender and origin to the test [22]. Therefore, the additional variable response times on the step size method can be applied to all the participants of the test without implications for gender and origin to the test.

III. RESEARCH METHOD

Items bank for the purposes of this research consisted of 600 items based on IRT 1 PL models. In this model, chances for somebody answered the item correctly depend only on the parameter level of difficulty items. Furthermore, two draft adaptive test developed is a conventional CAT (not randomized) and randomized. In this study the method of estimating the ability of test takers using MLE method and step-size.

In conventional CAT, the method of selecting the first item using medium difficulty level that starts with a range between -0.5 to 0.5 were selected randomly. Ability level estimation method using maximum likelihood estimation, but when the response answers the test taker is not yet patterned, estimating the level of ability of using a step size of 0.5. The next method of selecting items using the criteria of maximum information function. Items that have greatest value information function on specific capabilities have to be given to the test taker.

In the design of a randomized CAT, the design principle is the same with the conventional adaptive tests. The difference, at the election of the second point and so on using the principle of 5 – 4 – 3 – 2 – 1. The second item been selected randomly out of five (5) items which have the greatest information functions, the third item been selected randomly out of four (4) items which have the greatest information functions. The fourth item been selected randomly out of three (3) items which have the greatest information functions. The fifth item been selected randomly out of two (2) items which have the greatest information functions. Furthermore, for the sixth items and so the criteria for selecting the next items back to the maximum information function criteria that are not randomized (1 item).

To estimate the ability of the participants during the response of participants that not yet patterned, hence used the method step-size with an additional variable response time. For example, if there is no further information about the prior ability level of participants test, so the value $\theta_0 = 0$. Interval step size steadily increased of k (in this study was taken the value of k = 0.5). If the test participants responded with incorrectly answer, then estimate the ability of the participants into θ0 - k or 0 - 0.5 = -0.5. Meanwhile, when the the participants answered correctly, the estimated ability of the participants becomes θ0 + x k, or 0.5. x, where x is a positive constant multiplier and the amount depends on the category of the response time when the participants answered correctly. The procedure for estimating the level of participants ability with a step-size by a factor of response time participants are shown in Table 2.

Table 2. Estimated of Participants Ability Level in Step-Size Method Based on Response Time

Specification Notation: **time limit** = 150 second θ_0 = basic ability level = 0 **k** = step size = 0,5 **x** = constant multiplier $\theta_{ke-i} = \theta_{i-1} + xk$ (for the correct response) $= \theta_{i-1} - k$ (for the incorrect response)		**Response Corect Answer Consecutively**				**Response Incorect Answer Consecutively**			
		1st item	**2nd item**	**3rd item**		**1st item**	**2nd item**	**3rd item**	
		θ_1	θ_2	θ_3		θ_1	θ_2	θ_3	
Response Time Category	**Very Fast: x = 1.8** (Less than 30 s)	0, 9	1,8	2,7		-0,5	-1,0	-1,5	
	Fast: x = 1,6 (31 to 60 s)	0,8	1,6	2,4		-0,5	-1,0	-1,5	
	Medium: x = 1,4 (61 to 90 s)	0,7	1,4	2,1		-0,5	-1,0	-1,5	
	Slow: x = 1,2 (91 s.d. 120 s)	0,6	1,2	1,8		-0,5	-1,0	-1,5	
	Very Slow: x = 1 (121 to 150 s)	0,5	1,0	1,5		-0,5	-1,0	-1,5	
Extra time for 150 seconds. More than 300 s, the response is considered incorrectly.		0,5	1,0	1,5		-0,5	-1,0	-1,5	

In this study, the test termination criteria used were the test is stopped if the estimated value of the SEM has reached 0.30.

IV. RESULTS AND DISCUSSIONS

Summary of items bank statistics used in this study as follows:

Table 3. Statistics of Items Bank

General Description	Based on IRT with 1 PL
	The number of items = 600 items
Difficulty Level (b)	Minimum value = -3.0
	Maximum value = 3,0
	Number of items with medium difficulty level = 101 items

CAT test results showed that the number of items with a medium difficulty level, between -0.5 to +0.5 totaled 101 items. This case means that the possibility of the first items that appear to have the possibility a number of 101 items were taken randomly. When the participants answered correctly, then the second item to be displayed is item with the maximum information for θ = 0.5, and when the participants answered the item incorrectly, then the second item were to be displayed is items with the maximum information for θ = - 0.5. So it can be ensured that the conventional CAT, the second item consists only of the possibility one of the two items. In this study, the second item that appear are No.ID 577 (if answered correct) and No.ID 405 (if the answer is incorrect). Often the appearance of numbers No.ID 577 and No.ID 405 made a test CAT become unsafe due to the familiar questions.

The other case of item exposure that often appear is when using the step-size method. If the participants answered the questions always correct then the items appear are the items that have the maximum information value for θ = 0.5, 1.0, 1.5, 2.0, and 2.5, ie the second item with No.ID 557, the third item with No.ID 121, the fourth item four with No.ID 105, the fifth item with No,ID 247, and the sixth item with with No.ID 255. But, if the participants answer is always incorrect, the items appear is a matter that has maximum information value for θ = -0.5, -1.0, -1.5, -2.0, and -2.5, ie the second item with No.ID 557, the third item with No.ID 405, the fourth item with No.ID 125, the fifth item with No.ID 204, and the sixth items with No.ID 129.

If the participant's answer responses has been patterned (response answers already are correct and incorrect answers), then the next appeared item was quite varied because the first items that appears has a variable items is relatively large (101 items). However, with the use of maximum information function model to find items that match the level of the test participants' ability estimation, it is possible a lot of items that can not be displayed because they never get the maximum value for all levels of ability.

One proposed solution is to use a step-size method is based on the response time of participants answered correctly. Response time were stratified into groups based on the speed of response of participants correctly answered items raised by CAT. In the method step-size based on the response time, the formulation of the magnitude of the step-size value given additional constant multiplier based on the response time. The faster students respond to answers correctly then the bigger the multiplier constants.

Another proposed solution is to randomizes the maximum value of the function information. When the conventional CAT determines the items appear based on the maximum value of the function information (single), the CAT model of randomisation determines items appear by randomizing the maximum information function based group 5 - 4 - 3 – 2 - 1.

The second items obtained from randomize 5 greatest value function information based on the premise that in the early stages of the estimated level of proficiency test participants still contains an error value (SEM) high, so that not affect the result estimates the level of ability of the participants. Along with the many steps to estimate the ability of participants, the group randomized increasingly scaled down (to 4 - 3 - 2-1) along with decreasing error estimate. Thus, the items appear still refer to the estimated rate of the test participants' ability and does not affect the length of the test.

As an example will be given some comparative results of conventional CAT (which is not randomized) and the randomization CAT, as follows:

Table 4. Example of Conventional Cat (Not Randomized)

ITEM	1st	2nd	3th	4th	5th	6th	7th	etc
No. ID	284	**577**	**121**	**105**	**247**	430	283	etc
Response	***Correct***	***Correct***	***Correct***	***Correct***	***Incorrect***	***Correct***		etc
No. ID	25	**577**	**121**	**105**	357	126	92	etc
Response	***Correct***	***Correct***	***Correct***	***Incorrect***	***Incorrect***	***Correct***		etc
No. ID	146	**577**	**121**	518	139	450	77	etc
Response	***Correct***	***Correct***	***Incorrect***	***Incorrect***	***Incorrect***	***Correct***		etc

Table 5. Examples Of Randomization Cat

Item	**NO.ID**	**Response Answer**			**Explanation**
1	I.003 (b = 0.1)	Correct			First item is selected based on the level of medium difficulty (-0.5 ≤ b ≤ 0.5)
Selection Process of Second Item					
Because CAT is not patterned, then the process of selecting second items, using the method of step-size					
value of Ɵ that may appear	**Ɵ Selected**	**Alternative 5 Greatest Value of IIF**			
		IIF	**No. ID**	**No. ID Selected**	No.ID Selected appears by selecting randomly from 5 greatest value of IIF
0, 9	0.7	0,7225	I-068 (b = 0.7)	**I-170**	
0,8		0,722447802	I-170 (b = 0.69)		
0.7		0,722447802	I-480 (b = 0.71)		
0,6		0,722291238	I-034 (b = 0.68)		
0,5		0,722291238	I-406 (b = 0.72)		
Item	**NO.ID**	**Response Answer**			**Explanation**
2	I-170 (b = 0.69)	Correct			Second Item is selected based on the results of randomization
Selection Process of Third Item					
value of Ɵ that may appear	**Ɵ Selected**	**Alternative 4 Greatest Value of IIF**			
		IIF	**No. ID**	**No. ID Selected**	No.ID Selected appears by selecting randomly from 4 greatest value of IIF
1.6	1.2	0,7225	I-156 (b = 1.2)	**I-122**	
1.5		0,722447802	I-407 (b = 1.19)		
1.4		0,722447802	I-122 (b = 1.21)		
1,3		0,722291238	i-215 (b = 1.18)		
1.2					
Item	**NO.ID**	**Response Answer**			**Explanation**
3	I-122 (b = 1.21)	Correct			Third Item is selected based on the results of randomization
Selection Process of Fourth Item					
value of Ɵ that may appear	**Ɵ Selected**	**Alternative 4 Greatest Value of IIF**			
		IIF	**No. ID**	**N0. ID Selected**	No.ID Selected appears by selecting randomly from 3

2.1	1.7	0,7225	I-167 (b = 1,70)	**I-085**	greatest value of IIF
2.0		0,722447802	I-085 (b = 1,69)		
1.9		0,722447802	I-499 (b = 1,71)		
1,8					
1.7					
Item	**NO.ID**	**Response Answer**			**Explanation**
4	I-085 (b = 1,69)	Incorrect			Fourth Item is selected based on the results of randomization
Selection Process of Fifth Item					
Because CAT is already patterned, then the process of selecting fifth item, using the method of MLE					
Value of Ɵ	**IIF**	**No. ID**			**N0. ID Selected**
1.71	0,72250000	I.499 (b = 1.71)			**I .165**
	0,72244780	I.165 (b = 1.70)			
Item	**NO.ID**	**Response Answer**			**Explanation**
5	I-065 (b = 1,70)	Incorrect			Fifth item selected appears by selecting randomly from 2 greatest value of IIF
Selection Process of Sixth Item					
Value of Ɵ	**IIF**	**No. ID**			**N0. ID Selected**
1.40	0,72250000	I.279 (b = 1.40)			**I.279**
Selection Process Sixth item and Subsequent : The sixth item and subsequent items are selected based on the value of the largest IIF					

From the results of Table 3, it looks that CAT were not randomized appeared several items with the same identity, especially in the earlier pattern of the use of CAT. Meanwhile, in Table 4, the randomized CAT, many variations of possibilities items appear although the participant answers the same pattern. With so many variations items appear in the CAT randomized, it can reduce the level of item exposure, so that would make CAT more secure. Variations items appear on the actual randomized CAT has a difficulty level that is not much different from the CAT that were not randomized so it will not affect the length of the CAT test.

V. CONCLUSIONS AND RECOMMENDATIONS

A. Conclusions

From the analysis and discussion above, it can be concluded that the CAT by the method of randomization maximum information function with the criteria 5-4 - 3 - 2-1 by applying Maximum Likeli-hood Estimation (MLE) and the step-size based on the response time to estimate the the ability of participants can bring items with more variety. On the same participant answer pattern, the randomized CAT has item variation that have difficulty level similar to non-randomized CAT. Thus, it would produce a reduction in the level of exposure to the CAT items and increase security without reducing the CAT efficiency.

B. Recommendations

From this study, the design of randomized algorithms CAT is recommended to be applied to the adaptive test algorithms. CAT randomized algorithm does not reduce the level of efficiency and precision measurement, items on the initial order granted to the participants the test more varied so as to improve the safety test. This study uses a model 1 PL, it is recommended to use the model 2PL or 3PL to better examine variations items appear in the CAT.

REFERENCES

[1] F.M. Lord, “Applications of item response theory to practical testing problems”, Hillsdale, NJ : Lawrence Erlbaum Associates. 1980.

[2] H. Wainer, “Computerized adaptive testing: A primer, “ 2nd ed., Hillsdale, NJ: Lawrence Erlbaum Associates, 1990

[3] J.Q. Tian, D.M. Miao and X. Zhu, “An Introduction to the Computerized Adaptive Testing. US-China Education Review”, 2007,1, pp. 72-81.

[4] J.R. McBride and J.T. Martin, “Reliability and validity of adaptive ability tests in a military setting”. in D.J. Weiss, (Ed), New Horizons in Testing, New York, NY: Academic Press,1983, pp.223 – 236.

[5] D.R. Eignor, M.L. Stocking, M.L. and W.D. Way, “ Case studies in computer adaptive test design through simulation”, Research Report, Princeton, NJ: Educational Testing Service, 1993, pp. 93 – 96.

[6] D. Thissen and R.J. Mislevy, “Testing algorithms”, dalam H. Wainer (Ed.), Computerized Adaptive testing : A Primer, 2nd ed., Hillsdale, NJ: Lawrence Erlbaum Associates, 1990, pp. 103-135.

[7] W.P. Vispoel, “ Creating computerized adaptive test of music aptitude: Problem, solutions, and future directions,” in F. Drasgow and J.B. Olson-Buchanan (Eds.), Innovation in Computerized Assessment Mahwah, NJ: Lawrence Erlbaum Associates Publishers, 1999, pp. 151-176.

[8] C.N. Mills, “Development and introduction of a computer adaptive graduate record examination general test,” in F. Drasgow & J.B. Olson-Buchanan (Eds.), Innovation in Computerized assessment, Mahwah, NJ: Lawrencw Erlbaum Associates Publishers. Mills, 1999, pp. 123.

[9] A. Birnbaum, “Some latent trait models and their use in infering an examinee's ability,” MA: Addison-Wesley, 1986.

[10] F.B. Baker, “ Item response theory: Parameter estimation techniques,” New York: Marcel Dekker, Inc., 1992.

[11] B.G. Dodd, “The effect of item selection procedure and step-size on computerized adaptive attitude measurement using the rating scale model,” Applied Psychological Measurement, 4, 1990, pp. 355 – 366.

[12] D.J. Weiss, “Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education,” Measurement and Evaluation in Counseling and Development, 37, 2004, pp. 70-84.

[13] F.M. Lord, “A broad-range tailored test of verval ability,” Applied Psychological Measurement, 1, 1977, pp. 95-100.

[14] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, “Fundamentals of item response theory,” Newbury Park, CA: Sage Publication, Inc., 1991.

[15] K. Hambleton and H. Swaminathan, “ Item response theory,” Boston, MA: Kluwer Inc., 1985, pp. 94.

[16] D. Thissen, “Reliability and measurement precision,” in H. Wainer (Ed.), Computerized adaptive testing: A Primer (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum Associates, 1990, pp. 103-135.

[17] P.A. Dunkel, “Considerations in developing and using computer adaptive tests to asses second language proficiency” [Version Electronic]. Washington DC : Eric Clearinghouse On Languages And Linguistics Center For Applied Linguistics, 1999, http://www.cal.org/resources/digest/cat.html.

[18] W.J. Van der Linden, D.J. Scram and D.L. Schnipke, “Using response-time constraints in item selection to control for differential speededness in computerized adaptive testing” [Elektronic Version], LSAC Research Report Series, Newton, PA Law School Admission Council, 2003, http://www.lsac.org/lsacresources/Research/CT/CT-98-03.pdf .

[19] Lidia Martinez, “Time usage and candidate performance” [Electronic version], Chicago: Measurement Research Associates, Inc., 2009, http://www.rasch.org/mra/mra-06-09.htm

[20] Phil Higgins, “Candidate measured ability and use of time”, [Electronik version], Chicago: Measurement Research Associates, Inc. 2009, http://www.rasch.org/mra/mra-10-09.htm.

[21] Phil Higgins, “Item difficulty and time usage”, [Elektronik Version], Chicago: Measurement Research Associates, Inc., 2009, http://www.rasch.org/mra/mra-05-09.htm

[22] S.R. Chang, B.S. Plake and A.A. Ferdous, “Response times for correct and incorrect item responses on computerized adaptive testing”, 2005, Paper presented at the 2005 annual meeting of the America Education Research Association (AERA), Montreal, Canada.