ประสิทธิภาพของการทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์ โดยประยุกต์ใช้โมเดลการตอบสนองข้อสอบแบบพหุมิติ*

สุชาติ หอมจันทร์[1]

สมบัติ ท้ายเรือคำ[2] *บังอร กุมพล*[3]

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการทดสอบแบบปรับเหมาะ ด้วยคอมพิวเตอร์แบบพหุมิติ ที่กำหนดตัวแปรขนาดคลังข้อสอบ (A) จำนวน 100 ข้อ (a1) 150 ข้อ (a2) และ 200 ข้อ (a3) และตัวแปรเกณฑ์ยุติการสอบ (B) 3 เกณฑ์ คือ จำนวน 20 ข้อ (b1) SE(θ) $\leq$ .30(Reliability = .91) (b2) และ SE(θ) $\leq$ .43 (Reliability = .81) (b3) ข้อมูลที่ใช้เป็นการจำลองสถานการณ์ ด้วยโปรแกรม MATLAB โดยจำลองผู้สอบในแต่ละเงื่อนไข 1,000 คน วิเคราะห์ข้อมูลด้วยการวิเคราะห์ความแปรปรวนพหุคูณสองทาง (Two-way MANOVA) และทดสอบภายหลังด้วย Holtelling T^2 ผลการวิจัยพบว่ามีปฏิสัมพันธ์ (IAB) อย่างมีนัยสำคัญทางสถิติที่ระดับ .01 ระหว่างขนาดคลังข้อสอบ กับเกณฑ์ยุติการสอบ จึงทำการเปรียบเทียบรายคู่ (simple main effect) พบว่า

1. กรณีคลังข้อสอบ 100 ข้อ จะให้ความถูกต้องในการประมาณค่าความสามรถของผู้สอบได้ดีที่สุด เมื่อกำหนด SE(θ) $\leq$ .30 (b2) แต่จะให้ความแม่นยำในการประมาณค่าความสามรถของผู้สอบได้ดีที่สุด เมื่อกำหนด SE(θ) $\leq$ .43 (b3) เป็นเกณฑ์ยุติการสอบ

2. กรณีคลังข้อสอบ 150 ข้อ จะให้ความแม่นยำ และความถูกต้องในการประมาณค่าความสามารถของผู้สอบ ได้ดีที่สุด เมื่อกำหนดข้อสอบ 20 ข้อ(b1) เป็นเกณฑ์ยุติการสอบ

3. กรณีคลังข้อสอบ 200 ข้อ (a3) จะให้ความถูกต้องในการประมาณค่าความสามารถของผู้สอบได้ดีที่สุด เมื่อกำหนดเกณฑ์ยุติการสอบ (B) SE(θ) $\leq$ .30 (b2) แต่จะให้ความแม่นยำในการประมาณค่าความสามรถของผู้สอบได้ดีที่สุด เมื่อกำหนดข้อสอบ 20 ข้อ(b1) เป็นเกณฑ์ยุติการสอบ

คำสำคัญ: การทดสอบแบบปรับเหมาะด้วยคอมพิวเตอร์แบบพหุมิติ, ความเชื่อมั่น, เกณฑ์ยุติการสอบ

* วิทยานิพนธ์หลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัยและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยมหาสารคาม

[1] นักศึกษาหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาวิจัยและประเมินผลการศึกษา คณะศึกษาศาสตร์ มหาวิทยาลัยมหาสารคาม, E-mail: Suchart.hom007@gmail.com

[2] รองศาสตราจารย์, คณะศึกษาศาสตร์ มหาวิทยาลัยมหาสารคาม

[3] ผู้ช่วยศาสตราจารย์, ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม

The Efficiency of a Computerized Adaptive Testing (CAT) by Applying Multidimensional Item Response Models*

Suchart Homjan [1]

Sombat Thayreakum [2] *Bungon Kumpon* [3]

Abstract

The purpose of this research was to compare the efficiency of a Multidimensional Computerized Adaptive Testing (MCAT) by setting the variables for item pool (A) for 100 items (a1), for 150 items (a2), for 200 items (a3), and the variables of termination criteria (B) 3 criterion were 20 items (b1), SE (θ) $\leq$ .30 (Reliability = .91) (b2), and SE(θ) $\leq$ .43 (Reliability = .81) (b3). The data for simulation used the program of MATLAB by adapting each condition for 1,000 examinees. Data were Multiple analysis of variance (Two – way MANOVA) and Post Hoc Procedure by Holtelling T^2. The research result found, the interaction (IAB) was different at the statistical significance level of 0.01 between the item pool and the termination criteria so in comparing for simple main effect, it was found that

1. Case item pool 100 items: accuracy of the estimated ability of the examinees will be the best when setting the termination criteria for SE (θ) $\leq$ .30 (b2) but to the precision of the estimated ability of the examinees will be the best when setting the termination criteria for SE(θ) $\leq$ .43 (b3).

2. Case item pool 150 items: accuracy and precision of the estimated ability of the examinees will be the best when setting the termination criteria for 20 items (b1).

3. Case item pool 200 items: accuracy of the estimated ability of the examinees will be the best when setting the termination criteria for SE(θ) $\leq$ .30 (b2) but the precision of the estimated ability of the examinees will be the best when setting the termination criteria for 20 items (b1).

Keywords: Multidimensional Computerized Adaptive Testing, Reliability, Stopping Rule

* Research Article from thesis for the Doctor of Philosophy Degree, Educational Research and Evaluation Program, MahasarakhamUniversity, Thailand, 2015

[1] Student in Doctor of Philosophy Degree, Educational Research and Evaluation Program, Mahasarakham University, E-mail: Suchart.hom007@gmail.com

[2] Associate Professor, Department of Research and Educational Development Faculty of Education Mahasarakham University

[3] Assistance Professor, Department of Mathematics Faculty of Science Mahasarakham University

Introduction

Computerized Adaptive Testing (CAT) which is a form of computer-based test that adapts to the examinee's ability level. Computerized Adaptive Testing (CAT) has been widely used as it has multiple advantages over standard paper-and-pencil assessment tools. (Fliege et al., 2005) and can be used with the Item Response Theory (Weiss & Yoes, 1991). Multidimensional Computer Adaptive Testing (MCAT) expands the idea of Computerized Adaptive Testing. Besides MCAT could reduce the numbers of testing by CAT by about 30-50% and the numbers of traditional paper and pencil by about 70% without losing accuracy (Frey and Seitz, 2009, p.93). Item pool is very important in testing (Flaugher, 2000) and affects the performance of testing (B. Babcock and D. J. Weiss, 2009) The suitable item pool is between 100-200 items (Weiss, 1988).

Moreover, Termination Criterion affects the performance of testing (Nathan A. and Thomson, 2007) There are two kinds of Termination Criterion (Reckase, 2009, pp.335-336; Hambleton, R.K and H. Swaminathan, 1991): 1) Fixed-Length and 2) Variable-Length is a kind of termination criterion using the level of tolerance to accept by using the standard deviation of the estimate SE(θ) as a criteria to stop the exam. SE(θ) is the standard deviation of the distribution of the probability of the approximate value of the actual abilities (θ). The termination criterion with rule of variable length or the standard error of the estimate is correlated with reliability. The standard error of the estimate in Item Response Theory shows the result of standard error of ability estimation. While the classical test theory shows the reliability as the quality of the test which is related to the test score without any error. (Hambleton; Swaminathan; & Rogers, 1991, pp. 91-96). When substituting the reliability as .85, the standard error equals .385; standard deviation is 1. The standard error of the estimate is .385. (Ben Babcock and David J. Weiss, 2009) and When substituted the reliability at .91, the standard error equals .30 or The standard error of the estimate is .30. This value is commonly used as termination criterion in computerized adaptive testing. (Thissen, 1990, p.168). Reliability at .91 which is very high. One factor related to the high reliability is the numbers of items (Warm, 1978, p.77). On the other hand, the reliability in teaching in classroom is not as high as .90 (Lawrence M. Rudner and William D. Schafer, 2002, p.19) T.Dary Erwin showed that the reliability at .81 indicated the validity in measuring the homogeneity items. Connecting the reliability at .81 with computerized termination. The termination criterion is a representative of the whole test which has the standard error of estimate at .43.As long as the value of termination criterion is low, the more items will be used. What if the numbers of the items can be reduced and the performance of the test has no differences. What is the appropriate termination criterion that must be set? The purpose of this research is to compare the efficiency of Multidimensional Computerized Adaptive testing which used different item banks and different termination criterions.

Methodology

Multidimensional Computerized Adaptive Testing

Computerized adaptive testing composes of five components (Weiss & Kingsbury, 1984) and can be used in Multidimensional Computer Adaptive Testing (HANWOOK YOO. 2011, p. 30) as follow:

1. Model used in multi-dimensional response

The models used in multi-dimensional response starts with determine the inherent features p features as vector $\theta = \{\theta_1, \theta_2,\ldots, \theta_p\}$ when each p feature is from 1 item or more (Segall, 1996, p.333)

$$P_i(\theta) = P(U_i = 1|\theta) = c_i + \frac{1-c_i}{1+\exp[-Da_i^{'}(\theta - b_i 1)]}$$

When U_i is two random variable; i item response (U_i=1 when item i is correct and U_i = 0 when item i is incorrect), c_i is Guessing parameter of item i, b_i is Difficult parameter of item i, a_i is vector 1xp of discrimination parameter of item i, 1 is vector 1xp of 1s and D is Constant at 1.7

2. Ability Estimation

According to the Item response theory, There are two methods commonly used to estimate the true abilities of the examinees (θ) are Maximum Likelihood Estimation is a method of estimating the parameters of a statistical model and gives the estimated value that is closed to the actual abilities of the examinees.(Hambleton; & Swaminathan, 1991, pp.76-91) And Bayesian Estimation is the estimation suitable for computer adaptive testing. Bayesian Estimation includes Expected A Posteriori: EAP and Bayes Model: BM or Maximum A Posteriori: MAP. Bayesian Estimation use the information from a test response or likelihood function and prior information to determine the values of the estimation in ability of examinees. If the approximate value and the average distribution have very different ability value then the estimated ability value may be reduced to the preliminary average. (Meijer; & Nering, 1999) In this paper will show only Bayesian Estimation (Segall, 2000: 60 -64; Reckase, 2009, pp.144-145)

$$f(\theta/u)=\frac{L(u|\theta)\,f(\theta)}{f(u)}=\frac{L(u|\theta)\,f(\theta)}{\int_{-\infty}^{\infty}L(u|\theta)\,f(\theta)d\theta}$$

When L(u|θ) is probability function and f(θ) is the probability that occurred before θ

3. Item Selection

Item selection is to select the items from the item pool. So the item selection is very important in testing. The test will provide the maximum information when there is a difficulty parameter (b) nearby the examinees' abilities, high discrimination parameter (a), Guessing parameter (c) is near zero. When selecting the items, there must be several methods in selecting items. Each method can be used together with all the estimation methods and can also be used with several ways of selecting item. For example, item selection of Maximize Kullback-Leibler Information (Reckase, 2009, p.334) as follow.

$$K_i(\theta;\theta_0)=-E\left[\ln\frac{L(\theta_0|u_i)}{L(\theta_0|u_i)}\right]$$

When ln is natural logarithm and the function of L is

One item testing

$$K_i(\theta;\theta_0)=P_i(\theta_0)\ln\frac{P_i(\theta_0)}{P_i(\theta)}+Q_i(\theta_0)\ln\frac{Q(\theta_0)}{Q(\theta)}$$

More than one items testing

$$K_i(\theta;\theta_0)=\sum_{i=1}^{n}K_i(\theta,\theta_0)$$

4. Item pool

The researcher set the number of the item pool. There are 3 sizes including 100 items, 150 items and 200 items. (David J. Weiss, 1985) More item pools leads to the great performance. If there are 100 items or more in the item pool, will have good distribution. And it indicated that 150-200 items.

5. Termination Criteria

This study for each item pool, the same simulated response data were run 3 times, with each run using a different termination rule. The following were the termination conditions:

5.1 SE(θ) was below 0.30 (analogous to a reliability of 0.91) with a maximum of 100 items.

5.2 SE(θ) was below 0.43 (analogous to a reliability of 0.81) with a maximum of 100 items.

5.3 Fixed-length at 20 items with a maximum of 100 items.

5.4 SE(θ) was below 0.30 (analogous to a reliability of 0.91) with a maximum of 150 items.

5.5 SE(θ) was below 0.43 (analogous to a reliability of 0.81) with a maximum of 150 items.

5.6 Fixed-length at 20 items with a maximum of 150 items.

5.7 SE(θ) was below 0.30 (analogous to a reliability of 0.91) with a maximum of 200 items.

5.8 SE(θ) was below 0.43 (analogous to a reliability of 0.81) with a maximum of 200 items.

5.9 Fixed-length at 20 items with a maximum of 200 items.

Simulation

The data used in the research is from simulation by using MATLAB program of 1,000 examinees. The provided values are parameters of the examinees; the actual ability (θ). The test parameters in model of dimensional response are the response of multiple test is difficulty value (MDIFF), discrimination value (MDISC) guessing parameter (c) and the results of the responses (reply correctly get 1, not correctly get 0). In the simulation data, there is a replication in order to make validity and reliable information. Moreover, to find the value of actual ability in 50 rounds by sampling the distribution of normal curve have an average value that is equal to 0 and a standard deviation of 1 [$\theta \sim (0,1)$] in the range of -3.00 to 3.00 Parameter. The test difficulty parameter is in the range of -3.00 to 3.00.

Dependent variables

1. Root Mean Square Error: RMSE refers to the Precision of examinees' ability estimation which is the difference between the estimated ability and the mean square of actual ability or the variance of examinees' ability estimation. (Ben Babcock and David J. Weiss, 2009, p.9)

$$\text{RMSE} = \sqrt{\frac{\sum_{\text{j=1}}^{\text{N}} (\hat{\theta}_{\text{j}} - \theta_j)^2}{N}}$$

N is the number of examinees, $\hat{\theta}_{\text{j}}$ is the values of the examinees' ability estimation and θ_{j} is the value of the actual ability of the examinees

2. Average Bias refers to the accuracy of the estimated ability of the examinees. It is the average difference of estimated ability and the actual ability of the examinees. It can be the direction of the estimation including high or low. If the average bias equals 0 , it indicates that the examinees ability is accurate or no average bias. (Ben Babcock and David J. Weiss, 2009, p.9)

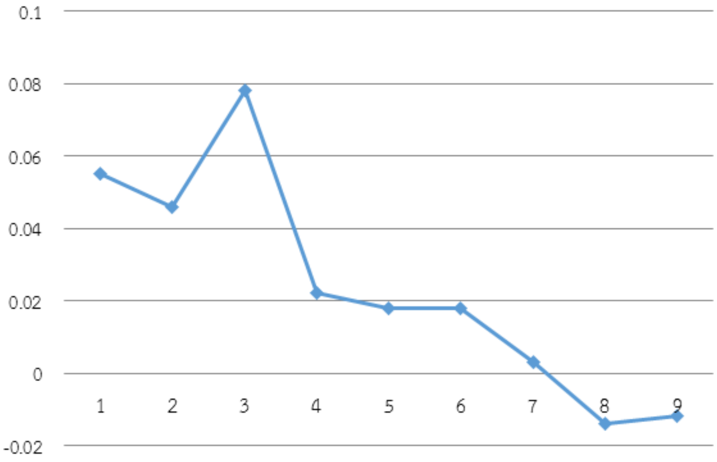
$$\mathrm{B}ias = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)}{N}$$

Results

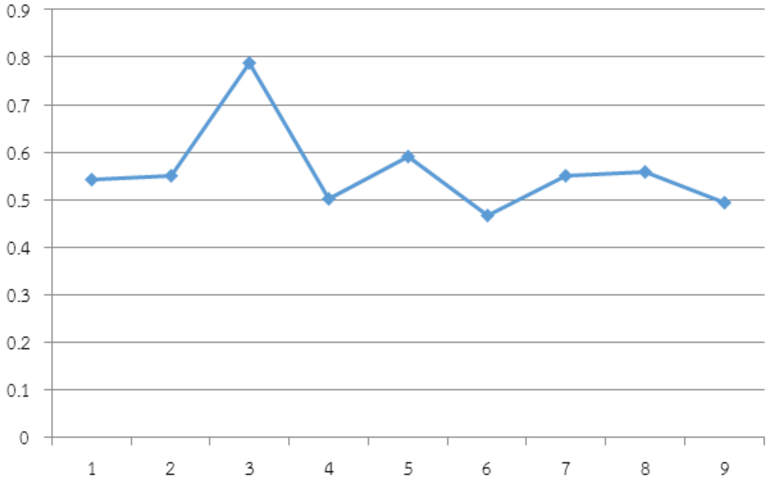
Table 1 shows numbers of estimation correctness and contestant contest, found that in a test 100 items had correctness estimation at SE(θ) $\leq$ 0.43 yet had accuracy estimation at SE(θ) $\leq$ 0.30. In a test 150 items had correctness and accuracy well providing 20 items in criterion. In a test 200 items had correctness estimation at SE(θ) $\leq$ 0.30 yet had accuracy estimation well in 20 items and indicates as the picture 1 and 2 below.

Table 1 The mean and standard deviation of Bias and RMSE categorized by the sizes of tests and termination of test criterion

method	Item	Stopping	Bias		RMSE	
	Pool	rule	$\overline{X}$	S.D.	$\overline{X}$	S.D.
1	100	.30	0.055	0.130	0.542	0.146
2		.43	0.046	0.150	0.551	0.148
3		20	0.078	0.178	0.788	0.117
4	150	.30	0.022	0.152	0.502	0.199
5		.43	0.031	0.250	0.590	0.212
6		20	0.018	0.075	0.467	0.159
7	200	.30	0.003	0.136	0.550	0.179
8		.43	-0.014	0.216	0.559	0.201
9		20	-0.012	0.065	0.494	0.151



Picture 1 The accuracy of estimated examinees' abilities from Bias



Picture 2 The accuracy of estimated examinees' abilities from RMSE

According to table 2, polynomial variance test of 2 independent variables Bias and RMSE, had sizes of pool tests and termination test criterions differently. The results showed that there was interaction significance at .01 level that affected on 2 variables having differently average. And the main factors found that the different sizes of tests and terminations of test criterions lead to 2 variables had differently average. From Levene's Test following table 3 below to test variance of 2 variables categorized by pool tests and termination test criterions differently, the results showed that there was interaction significance at .01 level. Consequently the researcher chose Post Hoc Tests or Multiple Comparison to compare the differences of average each group in table 5. It was a test of influenced between independent variable; the sizes of pool tests and termination test criterions found that Correlation Coefficient of 2 variables having a significance at .01 level. Thus pool tests and termination test criterions related to correctness and accuracy of estimated contestant's abilities from RMSE and Bias showing in picture 3 and 4 below.

Table 2 The pool tests and termination test criterions' interaction of Multidimensional computerized adaptive testing

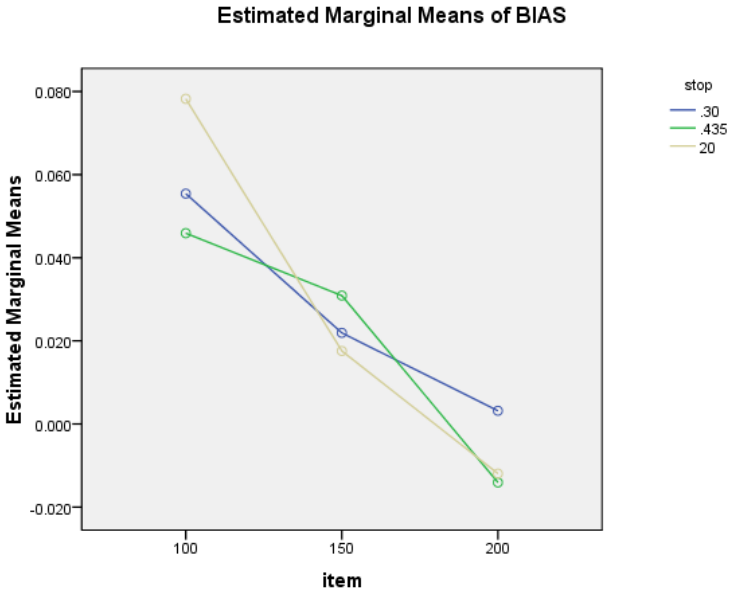
influences	Test statistics	value	F	Hypothesis df	Error df	sig
item	Pillai's Trace	.085	198.441	4.000	17982.000	.000
	Wilks' Lambda	.916	200.652a	4.000	17980.000	.000
	Hotelling's Trace	.090	202.865	4.000	17978.000	.000
	Roy's Largest Root	.077	347.816b	2.000	8991.000	.000
stop	Pillai's Trace	.022	50.421	4.000	17982.000	.000
	Wilks' Lambda	.978	50.683a	4.000	17980.000	.000
	Hotelling's Trace	.023	50.945	4.000	17978.000	.000
	Roy's Largest Root	.022	100.280b	2.000	8991.000	.000
item * stop	Pillai's Trace	.172	212.060	8.000	17982.000	.000
	Wilks' Lambda	.828	222.966a	8.000	17980.000	.000
	Hotelling's Trace	.208	233.918	8.000	17978.000	.000
	Roy's Largest Root	.208	467.149b	4.000	8991.000	.000

Table 3 The test of Bias and RMSE variance from Levene's Test

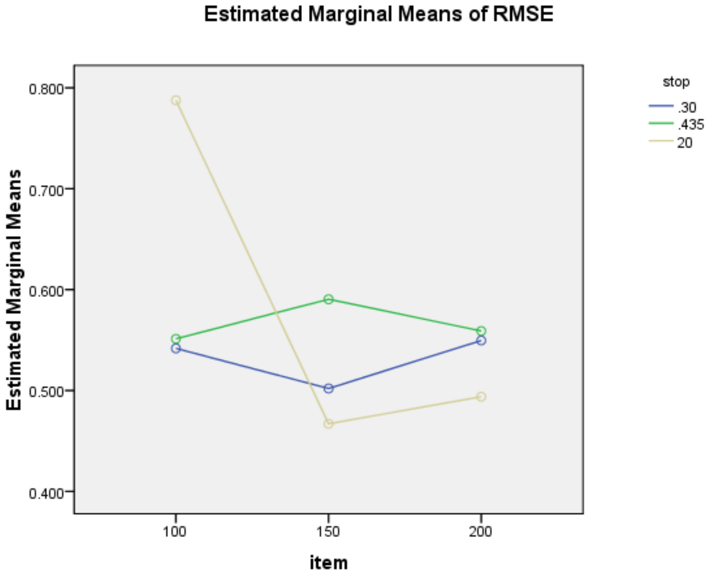
Independent variable	F	df1	df2	Sig.
Bias	130.095	8	8991	.000
RMSE	87.102	8	8991	.000

Table 4 The influence tests of independent variables

Variable sorts	Independent variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	BIAS	7.665a	8	.958	37.175	.000
	RMSE	69.651b	8	8.706	298.740	.000
Intercept	BIAS	5.730	1	5.730	222.306	.000
	RMSE	2825.159	1	2825.159	96938.723	.000
Item	BIAS	6.844	2	3.422	132.767	.000
	RMSE	20.272	2	10.136	347.785	.000
stop	BIAS	.086	2	.043	1.669	.189
	RMSE	4.211	2	2.105	72.239	.000
Item*stop	BIAS	.735	4	.184	7.132	.000
	RMSE	45.169	4	11.292	387.468	.000



Picture 3 The interaction of the sizes of pool test and Termination criteria test affecting on Bias



Picture 4 The interaction of the sizes of pool test and Termination criteria test affecting on RMSE

According to results and above pictures, found that independent variables has statistical significance, so the researcher provided a post test with Holtelling T^2, indicated that;

1. The comparison of correctness and accuracy in different pool tests

1.1 In case of 100 items in different termination criterion , found that the correctness and accuracy was different approximately significance at .01 level. When comparing by Pos Hoc found that Termination criteria; SE(θ) $\leq$ .30 20 items and SE(θ) $\leq$ .43 with 20 items, the correctness and accuracy of approximately value differently having significance at .01 level. And not different when comparing to termination criterion test SE(θ) $\leq$ .30 with SE(θ) $\leq$ .43 in significance at .01 level.

1.2 In case of 150 items in different termination criterion, found that the correctness and accuracy was not different estimated significance at .01 level. When comparing By Pos Hoc found that termination criterion; SE(θ) $\leq$ .30 20 items and SE(θ) $\leq$ .43 with 20 items, the correctness and accuracy of estimated value differently having significance at .01 level.

1.3 In case of 200 items pool in different termination criteria, found that the correctness and accuracy was different in statistical significance at .05 and .01 level respectively. When comparing the SE(θ) $\leq$ .30 and SE(θ) $\leq$ .43 by Pos Hoc found that the correctness of estimation had significant difference at .05 level but not different in accuracy. Considering the criterian at SE(θ) $\leq$ .30 with 20 items found that the correctness and accuracy of estimation had significant difference at .01 level. And the criterian at SE(θ) $\leq$ .43 with 20 items found that it was not different in correctness but significant different at .01 level in accuracy.

2. The comparison of correctness and accuracy in different termination criterion

2.1 In case of termination criteria SE(θ) $\leq$ .30 When comparing item pool between 100 item and 150 items and 150 items and 200 items found that the correctness and accuracy had the different significance at .05 level. For item pool between c 100 items and 200 items that the correctness was different in significance at .01 level but not accuracy.

2.2 In case of terminal criteria SE(θ) $\leq$ .43 when comparing between 100 items and 150 items found that that the correctness and accuracy being different significance at .01 level but not the correctness.. When comparing between 100 items and 200 items found that the correctness was different in significance at .01 level but not accuracy. Wand when using 150 items and 200 items found that the correctness and accuracy was different in significance at .01 level.

2.3 In case of termination criteria at 20 items When comparing between 100 items and 150 items, 100 items and 100 items and 150 items and 100 items found that the correctness and accuracy being different significance at .01 level.

Discussion

1. The correctness of estimated examinees' abilities

The correctness of estimated examinees' abilities categorized by item pool and termination criteria tests found that 200 items pool showing more stably corrective than 100 item pool and 150 items, especially termination criteria SE(θ) $\leq$ 0.30 indicating estimated correctness well. Considering quality of Ability Estimation can check at low correctness that meant approximated abilities was high depending on the sizes of accumulated tests. The more lots of pool test items, the more becoming suitable test with contestants. And the large pool test will be more informative than the small one. Moreover, parameter of the tests was possibly scattered depending with contestant's abilities, so a large pool test will be more efficient than a small. If defining termination criteria, the less error occurred, the less the differences of contestant's abilities. (Weiss, 1985) Weis stated (1985) that pool test with lots of items will be more efficient than a few pool test. Xing and Hambleton (2004) found that the more a pool test was large, the more became informative test, so there was also lots of informative tests. Related to Lord's (1980, pp.150-161) study, demonstrated by using a test with the 25 items length, the item pool with 363 items and 183 items. For 363 items found that it became more informative because a large size of the test was possibly suitable for contestant more than a small pool test. Above all, Way(1998) found that the size of pool test should be larger a test 6-8 times. When defining $SE(\theta) \leq 0.30$ affecting on estimation becoming more stable than any others depending the sizes of the test. If the pool test has influenced on estimation; pool test 100 items provided the less accuracy that the pool tests 150 and 200 items when using stable termination criterion with the 20 item length of the test will affect on the estimation of contestant. But comparing to termination criterion, $SE(\theta) \leq 0.30$ with $SE(\theta) \leq 0.43$. When using pool tests, 100 items, found that the number of items in termination criterion $SE(\theta) \leq 0.30$ was more than $SE(\theta) \leq 0.43$ because the less termination criterion, there would be more items and reliability. The errors of standard would be less respectively as correlation of formula; $S_E = \sqrt{1-R_{xx'}}$ (Warm, 1978, p.77) relating to Hambleton and Cook (1977) indicated that the errors of estimation in function correlation and informative function; $SEE = 1/\sqrt{I(\theta)}$. The value of informative function would be indicate a quality of estimation and it can be applies to find reliability in classical measurement instead. The research found that the correctness and accuracy show the values indifferently , so the termination criterion in reliability at .81 level being error standard .43, was a criterion of the number of test's items less than popular termination criterion nowadays but provided efficiency in tests indifferently

2. The accuracy of the examinees' Ability Estimation

The accuracy of the estimation of the examinees, according to the item pool and termination criterion, found that when using the termination criteria to fix the test at 20 items, the item pool of 150 items gives the best accuracy than those of the item pool of 100 and 200 items. That's because of the accuracy of the estimation is the nearest to 0. So the pool item of 150 items that used the termination criteria to fix the test at 20 items has the highest accuracy to estimate the examinees' abilities. Moreover, it is found that the item pool of 150 and 200 items that used the fixed termination criterion (20 items) give the similar accuracy of estimating examinees' ability. On the other hand, it is found that when using the pool item of 100 items and fixed termination criterion of 20 items, the accuracy in estimating examinees' ability is high. It showed that the sizes of pool tests, 100 items and stable termination criterion 20 items providing low in estimation of contestant's abilities because the length and sizes of the test can increase efficiency contestant's abilities depending on the sizes of the pool tests. If the length of the test was too much, it might become lass informative. And the test that too many items, it might be lack of equilibrium in pool test, so researcher should consider other factors relating designing a test adjusting with computerized adaptive testing. Seagall (2000) presented that the sizes of pool test should approximately 6 - 8 times of the length of the test. To define termination criterion was depended on researchers' condition that (Ben Babcock and David J. Weiss, 2009, p.18) presented that a termination criterion using standard error in estimation was more than one termination criterion, should take with stable item test. And from the research 15-20 items, it seemed low numbers suitably depending on precision in contestants' application.

References

Babcock, B., & Weiss, D. J., (2009). *Termination criteria in computerized adaptive tests: Variable-length CATs are not biased.* Preseted at the Realities of CAT Paper Session, June 2.

Erwin, T. D., (2000). *The NPEX sourcebook on assessment, volume 1: Definitions and assessment methods for critical thinking, problem solving, and Writing.* U.S. Department of Education, National Center for Education Statistics.

Flaugher, R. (2000). Item pools. In Wainer, H. (Ed.) *Computerized adaptive testing: A primer.* Mahwah, NJ: Erlbaum.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality Of Life Research*, *14*, 2277-2291.

Frey, A., & Seitz, N. N., (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation,* 35, 89-94.

Hambleton, R. K., & Cook, L.L., (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 14, 75-96.

Hambleton, R. K., & Swaminathan, H., (1991). *Fundamentals of Item Response Theory item response theory*. (2th ed). Newbury Park, C.A.: SAGE.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J., (1991). *Fundamentals of item response theory.* USA: SAGE.

HANWOOK YOO. (2011). *Evaluating several multidimensional adaptive testing procedures for diagnostic assessment*. Retrieved May 26, 2014, from ProQuest Dissertation & Theses databases. (Document ID: 3465252).

Lawrence, M. R., & William, D. S., (2002). *What teachers need to know about assessment.* National Education Association of the United States.

Lord, F. M., (1980). *Applications of item response: theory to practical testing problems*. New Jersey: Lawrence Erlbaum.

Meijer, R. R., & Nering, M.L., (1999). Computerized adaptive testing: Overview and introduction. *applied psychological Measurement*, 23(3), 187-194.

Nathan, A. T., (2007, January). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research & Evaluation.* (12), 1.

Reckase, M. D., (2009). *Multidimensional item response theory*. New York: Springer Science Business Media.

Segall, D. O., (1996). Multidimensional adaptive testing. *Psychometrika*, 61(2), 331-354.

_______. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden and C.A.W. Glas (eds.). *Computerized adaptive testing theory and practice (pp. 53-73).* Dordrecht: Kluwer Academic.

Thissen, D., (1990). *Reliability and measurement precision. in computerized adaptive testing: A primer*. by Howard Wainer, & et al. (pp.161-186). New Jersey: Lawrence Erlbaum.

Warm, T. A., (1978). A primer of item response theory. *Technical Report, 941278*. Oklahoma: U.S. Coast Guard Institute.

Way, W. D., (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *17*, 17-27.

Weiss, D. J., (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53(6), 774-789.

_______. (1988). Adaptive testing. *Educational Research Methodology and Measurement: International Handbook*. New York : Pergamon Press.

Weiss, D. J., & Kingsbury, G. G., (1984). Application of computerized testing to educational problems. *Journal of Educational Measurement*, 21(4), 361-375.

Weiss, D. J., & Yoes, M. E., (1991). Item response theory. In Hambleton, R.K., & Zaal, J. N., (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 69-95). Boston: Kluwer Academic.

Xing, D., & Hambleton, R. K., (2004). Impact of test design, Item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, 64(1), 5-21.

****