academicJournals

Vol. 8(13), pp. 1009-1021, 10 July, 2013
DOI: 10.5897/ERR2013.1133
ISSN 1990-3839 © 2013 Academic Journals
http://www.academicjournals.org/ERR

Educational Research and Reviews

Full Length Research Paper

Development of a computerized adaptive testing for diagnosing the cognitive process of grade 7 students in learning algebra, using multidimensional item response theory

Somprasong Senarat[1]*, Sombat Tayraukham[2], Chatsiri Piyapimonsit[3] and Sakesan Tongkhambanjong[4]

[1]College of Education, Roi-Et Rajabhat University, Thailand.
[2]Faculty of Education, Mahasarakham University, Thailand.
[3]Faculty of Education, Kasetsart University, Thailand.
[4]Faculty of Education, Burapha University, Thailand.

Accepted 16 May, 2013

The purpose of this research is to develop a multidimensional computerized adaptive test for diagnosing the cognitive process of grade 7 students in learning algebra by applying multidimensional item response theory. The research is divided into 4 steps: 1) the development of item bank of algebra, 2) the development of the multidimensional computerized adaptive testing program and a handbook for using the program, 3) the trial of the multidimensional computerized adaptive testing program and the handbook, and 4) the output evaluation of the developed process of the multidimensional computerized adaptive testing. The research and development output are as follows: The multidimensional computerized adaptive testing consists of the item bank. The item bank has criterion quality and it is divided into 2 parts: (1) item bank of order and graph, consisting of 59 items, (2) item bank of linear equation with one variable made up of 104 items. The multidimensional computerized adaptive testing was processed on Windows XP and Windows 7 and it could diagnose the cognitive process of grade 7 students learning algebra; it includes their ability to remember factual knowledge, to understand conceptual knowledge, to apply procedural knowledge, to analyze conceptual knowledge, and overall aid the classification of examinees into pass or fail categories. The trial of the multidimensional computerized adaptive testing program by the teachers and students indicated that they had satisfaction in processing the program. The evaluation of the multidimensional computerized adaptive testing by users indicated that it had a considerable efficiency in terms of utility, feasibility, propriety and accuracy.

Key words: Multidimensional computerized adaptive testing (MCAT), multidimensional item response theory (MIRT), diagnostic, cognitive process.

INTRODUCTION

In order to manage learning and teaching activities in order to acquire learning development or get students' feedback, teachers need to evaluate students continuously. The approach used for this was Diagnostic

*Corresponding author. E-mail: er2_somprasong@windowslive.com.

Assessment in which Cognitive Diagnostic Assessment was a learning and teaching model (Ketterlin-Geller and Yovanoff, 2009). It entails investigating the students' knowledge and skill process in learning to seek knowledge and understanding the strength and weakness of the learners (Leighton and Gierl, 2007). And educational evaluators believed that this model was based on the theory associated with the item response process. To implement the model in the educational assessment, psychological theory would be basically used (Rupp and Templin, 2008a). To find the psychological aspects, Computerized Adaptive Testing (CAT) was used for the examinees in order for them to be responsible to their abilities and to receive spontaneous feedback (Songsaeng, 2004; Kanjanawasri, 2007; Frey and Seitz, 2009). In future, there would be integration of testing between Modern Measurement Theories and Modern Technologies to respond to the need of information technology for making decision. This would make the measurement to be more precise and accurate (Kanjanawasri, 2007). Technology based on Modern Measurement Theories would help enhance the test making, test bank development, testing formatting, checking of test, analysis, testing result interpretation and testing report. This would make the testing system accurate precisely and flexibly (Kanjanawasri, 2007). This would relate to Multidimensional Computer Adaptive Testing (MCAT) by using Multidimensional Item Response Theory Models (MIRTM), the item response theory model consisting of various factors. Each variety would indicate the attributes used in Diagnostic Assessment (Sinhary et al., 2007; Haberman, 2008; Rupp and Templin, 2008b). Examinees would take the testing that one item gave more characteristics than the previous testing that one item gave only one attribute. Therefore MCAT would be more effective than CAT when testing with equal number of items - traditional paper and pencil. Besides this, 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). This could be processed through the computerized program and result and spontaneous feedback are given to the examinees faster.

Thinking process development or cognitive process is the aim of education related to learning process in learners' brain. This involves cognitive learning based on intellectual knowledge.

Thinking and problem solving

Significant intellectual perspectives accepted and used in current learning and teaching is of Bloom et al, (1956) and adapted by Anderson et al. (2001). So the researchers used the psychological measuring model - MIRT with 3 parameters in Algebra learning class for grade 7 students; they used processing dimensions and thoughts from Bloom et al (1956), newly adapted by Anderson et al. (2001) to apply the varieties in MIRT through 4 processes as follows: memorizing, understanding, knowledge application and analyzing.

According to research findings on problems of learning and teaching Mathematics, assessment theory, and learning assessment as mentioned above, the researcher was interested in MIRT development to diagnose the cognitive process in Algebra learning of grade 7 students. By this it was meant that the diagnosing approach to assess students' learning processes finds various processes and gives spontaneous feedback accurately and faster. This would be beneficial for teaching development and help reduce the time taken to diagnosis an individual. Apart from concluding the report and giving students spontaneous feedback, this could be more helpful for assessing and diagnosing the students in other aspects and so on.

Purpose

To develop a multidimensional computerized adaptive testing to diagnose the cognitive process in learning Algebra of grade 7 students by applying multidimensional item response theory.

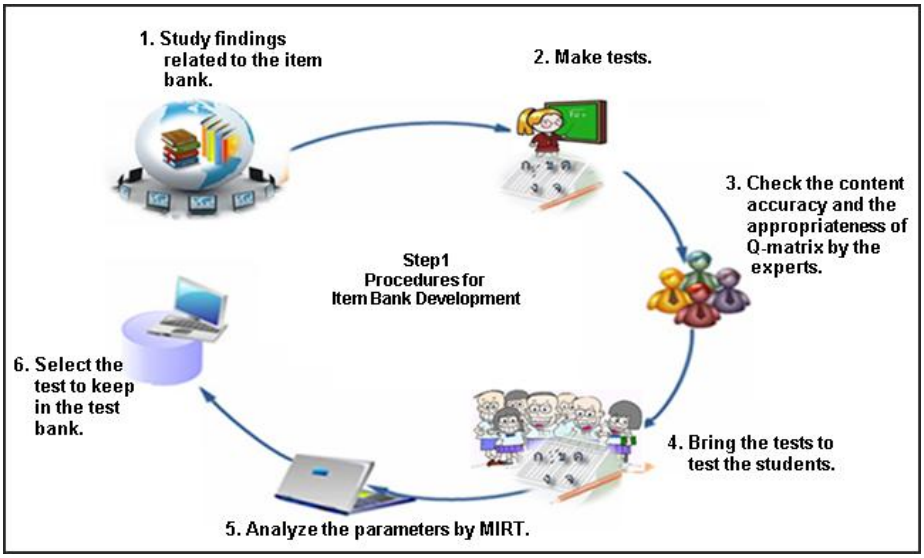
METHODOLOGY

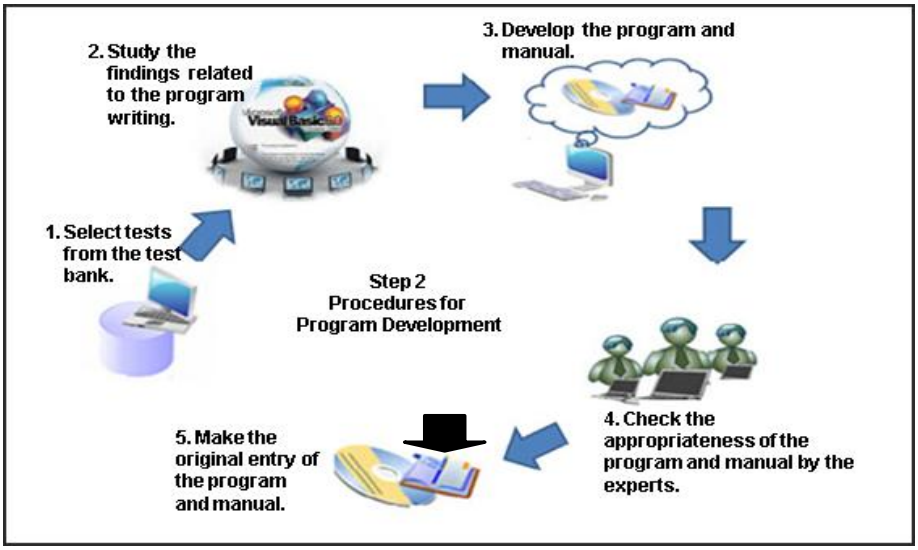
The methods were divided into 4 steps. In Figure 1, there are 4 procedures which consist of:

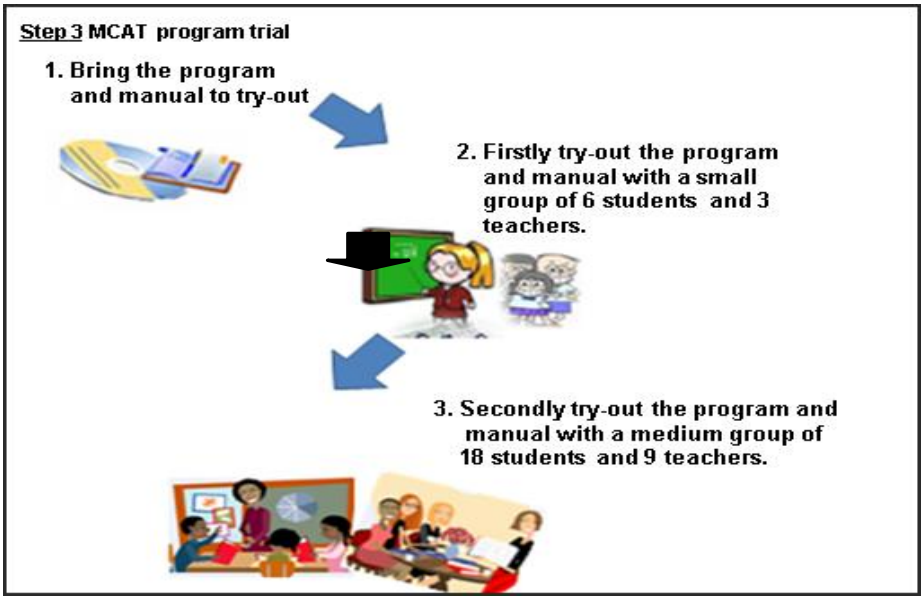
Step 1: The quality of item bank was developed by researcher, teachers and lecturers from a university who were teaching Mathematics; there were 423 items—Order and 140 graph items; linear equation with one variable include 283 items. Then the content validity and the appropriate relation between the items and cognitive processes of the items were inspected by 16 experts. There were 393 items approved including Order and 136 graph items in one testing set; and only 8 items of Order and Graph were congruently integrated. Meanwhile linear equations with one variable were 257 items and 7 items of these were also integrated. This item bank was brought to test 16,800 secondary students who were learning Order and Graph and Linear Equation with One Variable in the Northeastern area of Thailand in order to analyze the parameter.

Step 2: The quality checking process of MCAT program was developed and manual was used to take the item bank selected from Step 1 as a database in Visual Basic 6.0. The manual was created to inspect the appropriateness and accuracy of the computer program by 6 experts who were involved in computer programming or had academic positions and 1 expert who had doctorate degree in educational evaluation and assessment section.

Step 3: the program was taken from Step 2 into MCAT Testing Program Trial twice. Firstly the program and manual were used to try the small target group of 3 teachers, 3 students of average abilities and 3 students of low abilities in Mathematics to check if the basic program is working by using the structured interview on the appropriateness of the program. Secondly, the program and


1. Study findings related to the item bank.
2. Make tests.
3. Check the content accuracy and the appropriateness of Q-matrix by the experts.
4. Bring the tests to test the students.
5. Analyze the parameters by MIRT.
6. Select the test to keep in the test bank.
Step1
Procedures for
Item Bank Development


2. Study the findings related to the program writing.
3. Develop the program and manual.
1. Select tests from the test bank.
Step 2
Procedures for
Program Development
4. Check the appropriateness of the program and manual by the experts.
5. Make the original entry of the program and manual.


Step 3 MCAT program trial
1. Bring the program and manual to try-out
2. Firstly try-out the program and manual with a small group of 6 students and 3 teachers.
3. Secondly try-out the program and manual with a medium group of 18 students and 9 teachers.

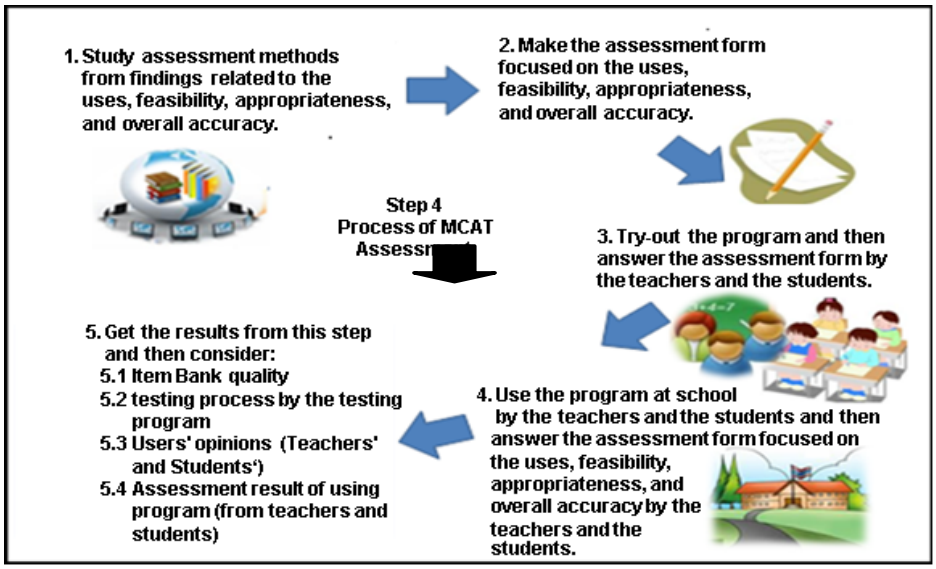


Figure 1. Procedures of developing an item bank by MCAT.

manual were used to try the medium target group of 6 teachers, 9 students of average abilities and 9 students of low abilities in Mathematics to check if the program is working by using the structured interview on the appropriateness of the program after adapting the users' comments in the first trial. Then the program and manual were improved for use in the ongoing procedures.

Step 4: MCAT processes were assessed by the researcher and the program adapted from (3) was used for 18 teachers and 174 students. The opinions about the uses, the possibility, the appropriateness, and the validity to cover all MCAT functions were tried and assessed. The teachers used the program to test their students in the classroom. The opinions of the teachers and the students were assessed after this trial.

MCAT development processes were detailed as follows:

Step 1: An item bank was developed as follows;

(1) Related documents were studied by analyzing theories, concepts, articles, researches and diagnosing multidimensional item response theory model of cognitive process and computerized adaptive testing.
(2) The researchers together with 4 math teachers (3 specialists and 1 math lecturer who have experience in teaching math in a university at least for 5 years) helped each other to do writing test of algebra for grade 7 students which included 5 choices based on cognitive process. There were 423 items—order; graph, 140 items; and linear equation with one variable, 283 items, which had a relation between the items and cognitive processes.
(3) With the content validity and Q-matrix checked by 16 experts, there were 393 items approved including order and graph (136 items) and linear equation with one variable (257 items). The 16 experts were 14 math lecturers (assistant and associate professors) who have been teaching math in the universities for at least 5 years and other 2 lecturers with doctorates degree in math who are experienced in evaluation and assessment in university. They were to consider each item on how appropriate the relations between the items and the cognitive processes were. Having the content validity and Q-matrix checked by 16 experts, there were 393 items approved: order and graph, 136 items; and linear equation with one variable, 257 items.
(4) The items were divided into 14 copies. Order and graph (4 copies): 40 items with 8 integrated items per copy, making a total of 136 items: linear equation with one variable (10 copies): 32 items with 7 integrated item per copy, making a total of 257 items.
(5) The complete test was used to test 16,800 secondary students in Northeastern area of Thailand who were learning Order and Graph and Linear Equation with One Variable.
(6) The students' test results were assessed by confirmatory factor analysis based on multidimensional item response model of multidimensional normal ogive model with NOHARM program. Then, the c value of each item was set at 0.20, while parameter, a value, discrimination power (a) and essiness intercept (d) were estimated from the possibility of the students' test ability in multidimensional normal ogive model (Bock and Schilling, 2003; McDonald, 1999; Samejima, 1974) as shown in equation 1,

$$P(\mu_{ij}=1|\theta_j, \mathbf{a}_j, c_i, d_i) = c_i + (1+c_i)\frac{1}{\sqrt{2\pi}}\int_{-z_i(\theta_j)}^{\infty} e^{\frac{-t^2}{2}}dt, \quad \textit{Where } z_i(\theta_j) = \mathbf{a}_j\theta_j' + d_i \quad (1)$$

Where $P(\mu_{ij}=1|\theta_j, \mathbf{a}_j, c_i, d_i)$ is the probability of a correct response for examining j on test item i and in m dimensional space, u_{ij} is the item response for person j on item i (1 correct; 0 wrong), $\mathbf{a}_j$ is a vector of parameter that specifies the discrimination power of the item i on each of the m dimension in the space, c_i is a parameter that specifies the probability of correct response for persons who are low on all of the dimensions, d_i is a parameter related to the difficulty of item i, (Essiness intercept) , θ_j is a vector of parameters that describe the location of person j in an n-dimensional space, and e is the mathematical constant 2.7182818.
(7) NOP (Non–orthogonal procrustes method) was applied to equate with the discrimination power and essiness intercept parameters as shown in equations 2 and 3 (Reckase and Martineau, 2004),

$$\mathbf{a}_i^* = \mathbf{a}_i'\mathbf{T} \quad (2)$$

$$d_i^* = d_i + \mathbf{a}_i^*\mathbf{T}\mathbf{m} \quad (3)$$

Where $\mathbf{a}_i^*$ and d_i^* are the values of parameters from the

comparison form transformed to match the metric of the base form, $\mathbf{a}_i$ is a vector of discrimination parameters; item i of the comparison form, d_i is a parameter related to item difficulty; item i of the comparison form, **m** is a translation vector for location, **T** is an orthogonal procrustes rotation matrix for positioning calculated from $\mathbf{T}=(\mathbf{A}'\mathbf{A})^{-1}\ \mathbf{A}'\mathbf{B}$ while **A** is the matrix of the discrimination power of the comparison form, **B** is a parameter matrix of the base test discrimination power of the base form.

(8) Multidimensional discrimination (MDISC) and Multidimensional difficulty (MDIFF) were inspected to meet the test quality as shown in equations 4 and 5 respectively (Reckase and McKinley, 1991; Reckase, 2009).

$$MDISC = \sqrt{\sum_{k=1}^{m} a_{ik}^2} \tag{4}$$

$$MDIFF = \frac{-d_i}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}} \tag{5}$$

(9) The test items were chosen to an item bank by considering how the test items fit the standard criteria, which are: the discrimination of each dimension should not be negative value, multidimensional difficulty should be between -4.00 to 4.00, and the discrimination should not be too different among the dimensions. And Microsoft Access 2003 was applied for managing database of an item bank.

Step 2: A multidimensional computerized adaptive testing (MCAT program) program was developed by Visual Basic 6.0 as follows;

(1) Study related documents, concepts, theories, articles and researches to develop the MCAT program. Then bring the item bank from step 1 as a database to develop the program.
(2) Design a structure and elements of the program--program screen including buttons for working control such as suggestions before testing, using handbook, choosing content, and exist buttons, and other formative designs.
(3). Make a working processes diagram of the MCAT program. The program would work orderly namely, log in, start, valuate the common ability of the test ($\theta=0$), choose the best test item, show the first item, answering result ; after that valuate the tester according to the standard and finally, terminate testing, report the result and end the test if it supports the criteria. The program would, on the other hand, adapt and choose the best further item if it does not meet the standard criteria. The processes are shown in Figure 2.

The details of the diagram of the MCAT program are;

Login: the program will login to the page that contains using suggestions of a program before testing; the tester has to choose the content, fill the personal information and save it before taking a test.

Start doing the first item which has the highest information value: ($\mathbf{I}_i$) is from the estimator named Largest Decrement in the Volume of the Bayesian Credibility Ellipsoid, while $\mathbf{I}_i$ is the maximum former information matrix,

$$\mathbf{w}_i = \mathrm{D}^2\mathbf{a}_i\mathbf{a}_i'\mathbf{w}_i^*, \quad \mathbf{w}_i^* = \left[\frac{\mathbf{q}_i(\boldsymbol{\theta})}{\mathbf{p}_i(\boldsymbol{\theta})}\right]\left[\frac{\mathbf{p}_i(\boldsymbol{\theta})-\mathbf{c}_i}{1-\mathbf{c}_i}\right]^2$$

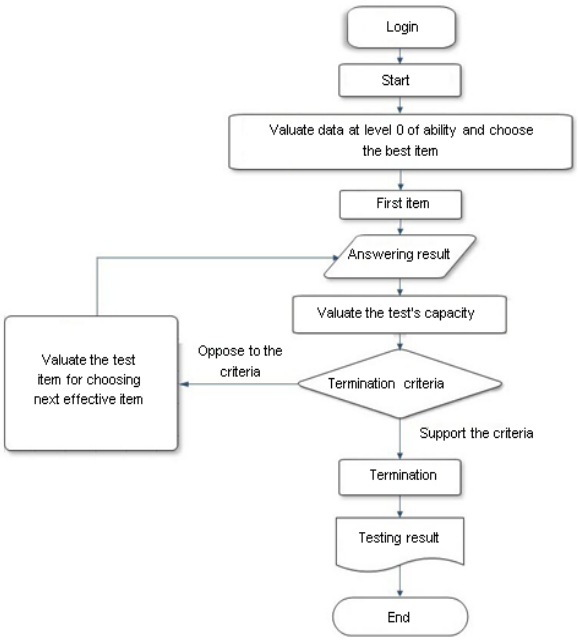


Figure 2. Diagram of the MCAT program.

$\boldsymbol{\varphi}^{-1}$ is invert value of variance covariance matrix as shown in equation 6 (Segall, 2010)

$$\mathbf{I}_i = \boldsymbol{\varphi}^{-1} + \mathbf{w}_i \tag{6}$$

Valuate the testers' ability to get a database in order to choose the next item by Fisher's information from capacity valuation of the Largest Decrement in the Volume of the Bayesian Credibility Ellipsoid which was presented by Segall (2010). The processes are;

The possibility of answering correct answer was calculated through Multidimensional Normal Ogive model, $\mathbf{z}_i(\boldsymbol{\theta}_j)$ is $d_i+\mathbf{a}_i\boldsymbol{\theta}_j=d_i+a_{i1}\theta_{j1}+a_{i2}\theta_{j2}+\ldots+a_{ip}\theta_{jp}$ as shown in equation 9, (Bock and Schilling, 2003; Reckase, 2009).

Estimate a tester's ability by Largest Decrement in the Volume of the Bayesian Credibility Ellipsoid. The posterior density $\mathrm{f}(\theta|u)$ contains all existing information about θ and is used as basis to provide point and interval estimates of ability parameters of θ, as shown in equation 7 (Reckase, 2009; Segall, 2010).

$$\mathrm{f}(\theta|u) = \frac{L(u|\theta)\mathrm{f}(\theta)}{f(u)} = \frac{L(u|\theta)\mathrm{f}(\theta)}{\int_{-\infty}^{\infty} L(u|\theta)\mathrm{f}(\theta)\mathrm{d}\theta} \tag{7}$$

where $L(u|\theta)$ is a likelihood function; $L(u|\Theta)=f(U_{i1}=u_{i1},U_{i2}=u_{i2},U_{in}=u_{in}|\theta)=L(u|\theta)=\prod_{i\in S_n}p_i(\theta)^{u_i}\,q_i(\theta)^{1-u_i}$

Where the product runs over the set of administered (or selected) items $s_n=\{i_1,i_2,\ldots i_n\}$, and $\mathrm{q}_i(\boldsymbol{\theta}) = 1-\mathrm{p}_i(\boldsymbol{\theta})$; $\mathrm{p}_i(\boldsymbol{\theta})$ is probability of answering a correct answer. The ability to express

$f\left(U_{i1}=u_{i1},U_{i2}=u_{i2},U_{in}=u_{in}|\theta\right)$ as a product of terms that depend on individual item-response functions that lead to computational simplifications in terms of selection and scoring, and $f(u)$ is the marginal probability of **u** given by,

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

Where $f(\theta)$ is the multivariate normal density function given by $f(\theta)=(2\pi)^{\frac{p}{2}}|\varphi|^{-\frac{1}{2}}exp\left[-\frac{1}{2}(\theta-\mu)\varphi^{-1}(\theta-\mu)\right]$ and $\frac{\partial}{\partial\theta}ln\ f(\theta|u)=0$ and $\frac{v_1=(p_i(\theta)-c_i)(u_i-p_i(\theta))}{(1-c_i)p_i(\theta)}$ and $J_s(\theta)=\frac{\partial^2}{\partial\theta\partial\theta'}ln\ f(\theta|u)=D^2\sum_{i\in S}a_i a_i' w_i-\varphi^{-1}$ and $w_i=\frac{q_i(\theta)[p_i(\theta)-c_i][c_i u_i-p_i^2(\theta)]}{p_i^2(\theta)(1-c_i)^2}$

$$\delta^{(m)}=\left[J_s(\theta^{(m)})\right]^{-1}\frac{\partial}{\partial\theta}ln\ f(\theta^m|u)$$

Selecting another test item that was suitable for the tester's ability by Decrement in the Volume of the Bayesian Credibility Ellipsoid Method of Segall (2010). The chosen item should have highest afterward information matrix; $I_{i|s_{k-1}}$ is shown at equation 8. An estimation of a temporality ability of $\hat{\theta}_k$, (when k=0) was done by specifying a set of capacity distribution average value as 0. The first step was to calculate the value of a co-variance invert matrix before (φ^{-1}); then find out $W_{s_{k-1}}$ as presented in equation 9 when W_j of item j was calculated from equation (17) and total value of matrix w that was from a former item. $\sum_{j\in s_{k-1}} W_j$ was calculated from all chosen items. Finally, find a value of matrix w which was from the next chosen test item by equation 10.

$$I_{i|s_{k-1}}=\varphi^{-1}+W_{s_{k-1}}+W_i \quad (8)$$

$$W_{s_{k-1}}=\sum_{j\in s_{k-1}} W_j \quad (9)$$

$$W_i=D^2 a_i a_i' w_i^* \quad (10)$$

The 2 criteria used for ending testing were standard error of the estimate value and the constant numbers of a test item. Standard error of the estimate of a tester's ability with $SE(\theta)$ is standard error of the estimate; θ and $I(\theta)$ were the test information given to any skilled person at θ (Kanjanawasri, 2007) that could be calculated as in equation (11),

$$SE(\theta)=\frac{1}{\sqrt{I(\theta)}} \quad (11)$$

The report of the personal tester's ability was divided by the contents and cognitive processes and presented through message and graph.

Visual Basic Version 6.0 of Windows XP was applied to develop MCAT.

Inspect the developed program for its effectiveness and improvement to meet the standard if some errors are found.

Make a handbook of using the program in order to use it correctly and orderly; it should include processes of installing, how to run and use main and minor screens and interpreting test result.

Let the developed program and its handbook be inspected by 6 computer experts

Step 3: Test using the Computerized Adaptive Testing Program two times;

First trial: Firstly the program and manual were used to test 3 teachers, 3 students of average abilities and 3 students of low abilities in Mathematics to check if the basic program is working by using the structured interview on the appropriateness of the program.

Second trial: Secondly, the program and manual were used to test 6 teachers, 9 students of average abilities and 9 students of low abilities in Mathematics. They were divided into 3 groups. The program and the manual made by the researchers were used to get the information to adapt and improve test on the issues. They include: the comfort to install program, interpretation of speed, appropriateness in presenting testing of the program, data saving, systematic working of the program, accessibility of the program, screen management, font and background of the text, the appropriate interaction between the users and the program, help functions in the program, the objectives of the program, the overall content coverage in program using, comprehensible language used in the manual, the font sizes in the manual, content stepping order in the manual, pictures and photos in the manual, stylish typing to interest the users and comprehension and accessibility. Then the program and the manual were adapted and improved after the users' comments in the first and second trials. The structured interview had been used as a tool to collect the data.

Step 4: Evaluating the processes of MCAT focused on;

1. The quality of the item bank and MCAT program.
2. Testing process of MCAT

The first two processes had been already conducted in the first to the third step.

Users' opinions from 192 users including 18 teachers and 174 students by studying the assessment methods from findings related to uses, possibilities, appropriateness and accuracy

Tools made for assessing teachers' and students' opinions using **MCAT program with 5 rating scales are toward 4 aspects:** (1) Utility: to assess the appropriateness of the MCAT program developed by the researchers to meet the users' need authentically (2) Feasibility: to assess the appropriateness of the MCAT program developed by the researchers to use in real life (3) Propriety: to assess the appropriateness of the MCAT program developed by the researchers to link with the educational goals, curriculum, and teaching and learning objectives without being against the policy, legislation and morality

Accuracy: this entails assessing the appropriateness of the MCAT program developed by the researchers to diagnose the learners accurately by covering the real situations and then the assessment forms made by the researchers were brought to the research

Table 1. Results of analysis of the chosen items' parameters in an item bank.

Contents	Statistics	Parameter						
		Dimension 1 (a_1)	Dimension 2 (a_2)	Dimension 3 (a_3)	Dimension 4 (a_4)	D	MDISC	MDIFF
Order and graph	$\overline{X}$	1.137	1.016	0.568	0.931	-0.258	1.744	0.188
	SD	1.261	0.875	0.688	0.527	0.973	1.282	0.831
	Min	0.067	0.111	0.054	0.041	-2.870	0.350	-1.663
	Max	5.207	4.208	2.113	1.326	2.114	5.296	3.133
Linear equations in one variable	$\overline{X}$	1.199	0.718	0.804	0.802	-1.876	1.750	1.058
	SD	0.915	0.545	0.564	0.573	1.974	0.996	0.875
	Min	0.027	0.022	0.022	0.056	-9.827	0.448	-0.925
	Max	4.289	2.662	2.599	2.279	0.494	4.807	3.875

adviser to check and give suggestions for adapting to appropriateness.

Assessment forms made by the researchers were submitted to 4 experts who were teaching Mathematics in universities and other 2 experts who have experience and were working with computer; 1 expert in evaluation and assessment, and 1 expert in psychology and guidance for checking IOC

Teachers and students were made to use the MCAT program. After the trial, the teachers and students suggested the use of the program from the assessment forms and open ended questionnaires.

RESULTS

The development of Algebra results of grade 7 students consisted of 2 parts: Order and Graph, and Linear Equation with one variable.

The results of choosing an item from an item bank showed that 59 items of Order and Graph and 104 items of Linear Equation with one variable were chosen. The parameters of the chosen Order and Graph items were; the discriminations of dimensions 1, 2, 3 and 4 were between 0.067 to 5.207, 0.111 - 4.208, 0.054 to 2.113, and 0.041 to 1.326 respectively. The average discrimination value in dimension 1 was the highest and followed by the ones in dimensions 2, 4 and 3; 1.137, 1.016, 0.931 and 0.568 respectively, while their standard deviations were 1.261, 0.875, 0.527 and 0.688 respectively. Its d value was between -2.870 to 2.114; average: -0.258 and SD was 0.973. MDISC was between 0.350 to 5.296; average: 1.744 and SD was 1.282. MDIFF was between -1.663 to 3.133; average: 0.188 and SD was 0.831. The parameters of the Linear Equation with one variable item were; the discriminations in dimensions 1, 2, 3, and 4 were between 0.027 to 4.289, 0.022 to 2.662, 0.022 to 2.599, and 0.056 to 2.279 respectively. The average discrimination ranged from the highest to the lowest one: 1, 3, 4, and 2: 1.199, 0.804, 0.802 and 0.718 and its standard deviations were 0.915, 0.564, 0.573 and 0.545 respectively. Its d value was between -9.827 to 0.494; average: -1.876 with 1.974 of its SD. MDISC of the chosen items was between 0.448 to 4.807; average: 1.750, with 0.996 of SD. And MDIFF was between -0.925 to 3.875; average: 1.058 with 0.875 of SD (Table 1).

According to the results of developing MCAT program, the researchers got the effective program with its handbook. The program could be applied for cognitive process diagnosis of grade 7 students. It could be used with Windows XP and Windows 7 and its elements are as follows.

Testing management parts including main and sub-screen of the whole test as shown in Figures 3 to 9.

Figure 3 shows the main screen including command button, advice before testing, handbook, choosing the contents and exiting the program

Figure 4 consists of 2 contents of Algebra: Order and Graph and Linear Equation with one variable.

Figure 5 shows that the tester would have to fill the personal information including name/number and the school's name. When clicking the Start button, the first item would be seen at a sub-screen as shown in Figure 6.

Figure 7 shows that students' answers were false after choosing and confirming the answer. In item 1, only 1 aspect was asked - ability to memorize fact. So when the students choose the wrong choice, it means that they lack good ability to memorize fact.

Figure 8 shows students' answers were true after choosing and confirming the answer in item 1. So when the students choose the right choice, it means that they have good ability to memorize fact.

Figure 9 consists of the second item. The program shows the next item after students had finished choosing the first item. Then the program ran continuously until item 15 was finished; also, the program would assess the students' ability as well.

The tester could see a report result shown in Figure 10. Figure 10 sums up the testing results by instructing button; 1) Analyzing results of the testers' cognitive ability 2) Summing up the individual result of each item and the

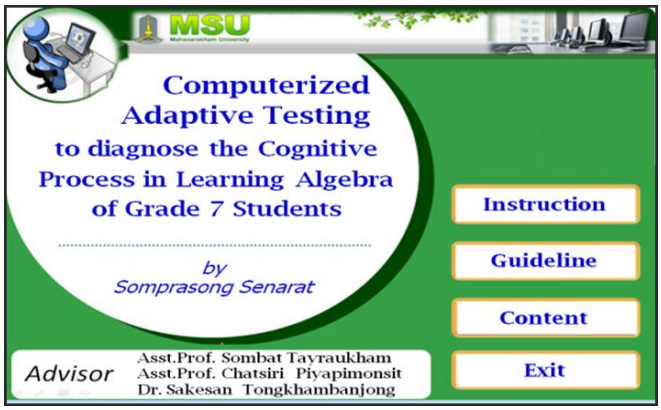


Figure 3. Main screen.

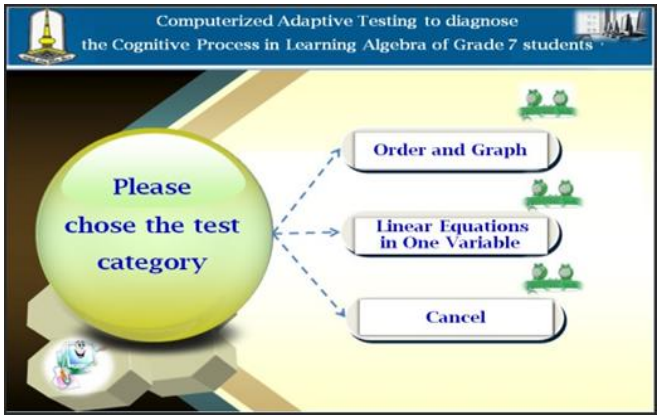


Figure 4. Choosing a content.

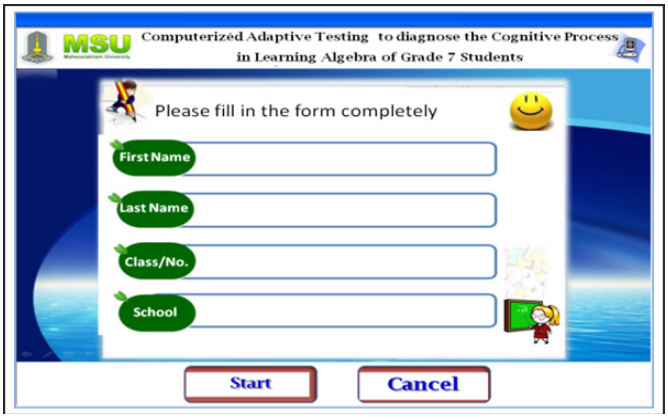


Figure 5. Fill out the personal information.

result of all items 3) Graph showing the cognitive competence of the testers 4) Information.

Figures 11 to 15 show the confidence and standard errors of the items.

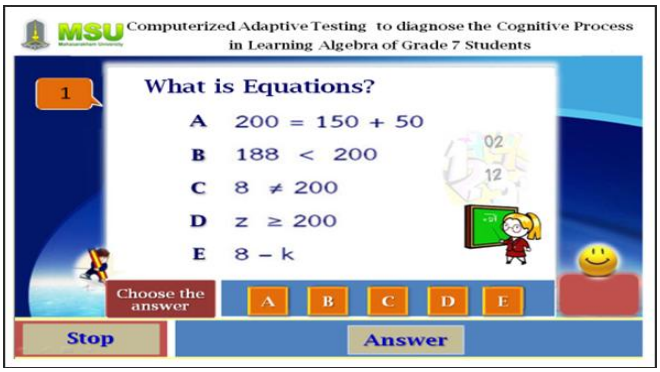


Figure 6. The first item.

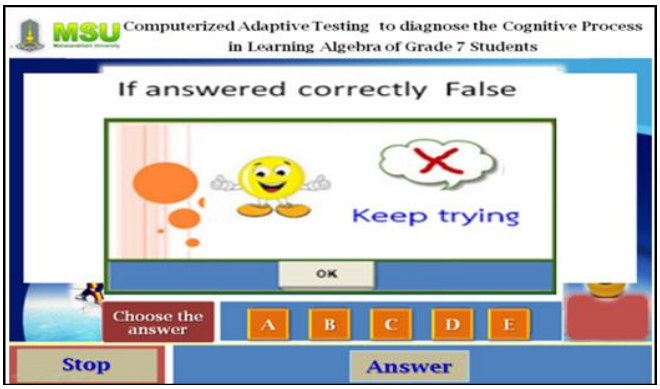


Figure 7. The result of testing; if answered falsely.

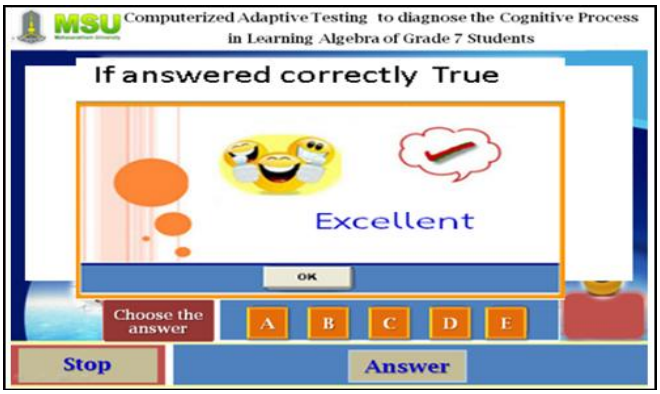


Figure 8. The result of testing; if answered correctly.

Figure 11 reports the testers' ability followed by the cognitive intellectual process. In this part, teachers and students were able to look at the report in every issue except the graph of the testers' ability followed by the cognitive intellectual process and information, the confidence and standard errors of the items in which the teachers would give more explanations.

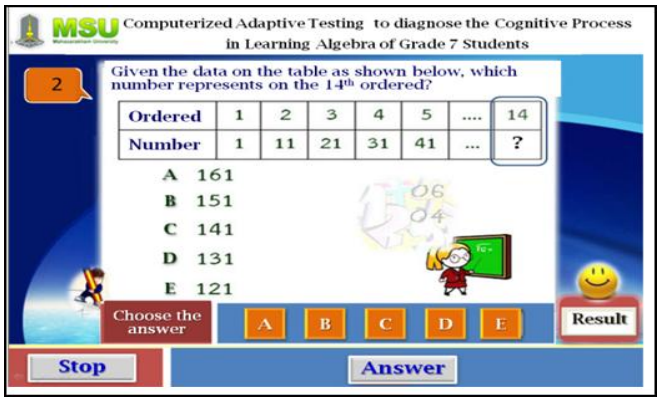


Figure 9. The second item and the next item.

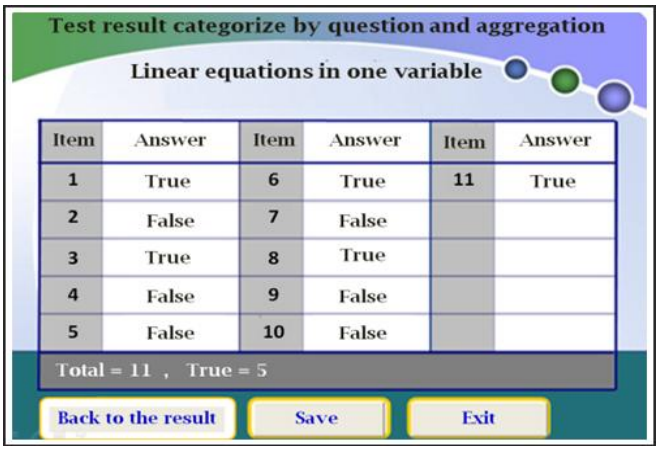


Figure 12. Results of answering.

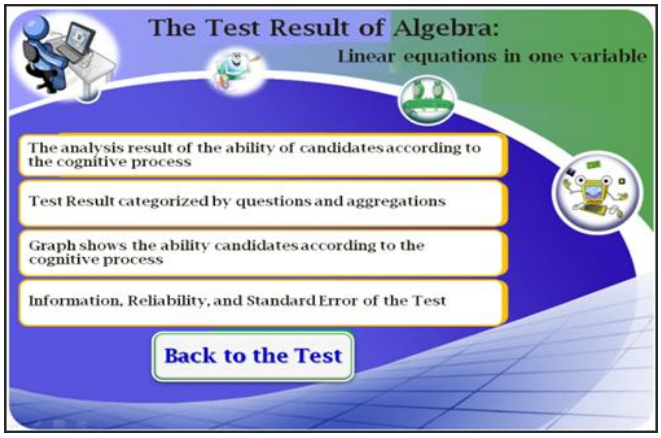


Figure 10. Result of answering.

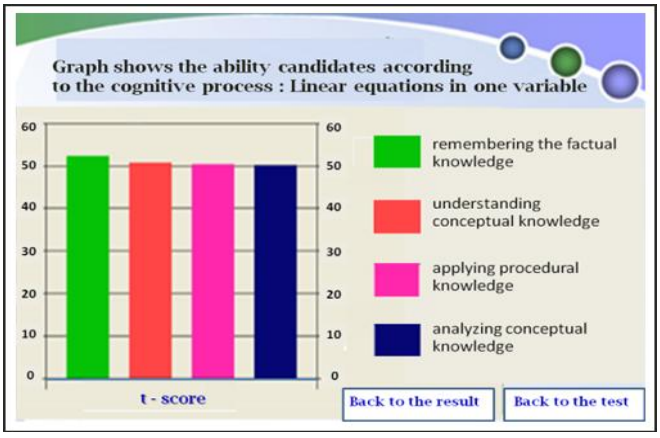


Figure 13. Graph of a tester's ability.

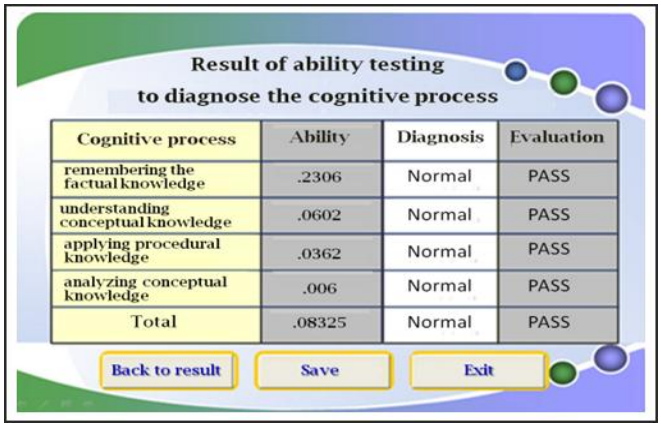


Figure 11. Analysis result of the testers' ability.

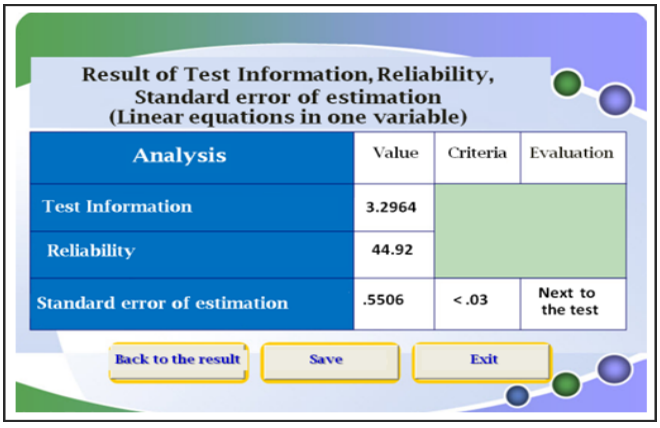


Figure 14. Information, reliability and standard errors.

Figure 12 reports the result conclusion of the item answers divided individually and overall by summarizing the number of all items taken by the students and the number of items chosen correctly.

Figure 13 shows the graph of the testers' ability followed by the cognitive intellectual process with t-score. In this part the students would get some more suggestions from the teachers.

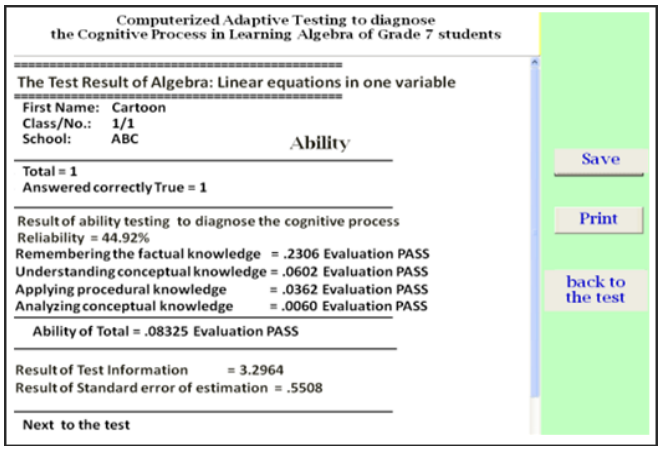


Figure 15. Saving and printing the test results.

Figure 14 presents the information, confidence, standard errors of the test, (If the testing does not oppose the ending testing criteria, the result report would appear automatically). In this part the students would get some more suggestions from the teachers.

Figure 15 saves the test result and printing. The program enabled the testers to record and to print out the result.

The results of appropriation and accuracy of MCAT and its handbook are presented in Tables 2 and 3.

The result of the accuracy and appropriateness of the MCAT from the assessment answered the opinions on the program given by 5 experts in computer and 1 expert in evaluation and assessment. Overall, the experts viewed the program using highly agreed level rating with an average point of 4.21 and standard deviation of 0.73; the average value was between 3.83 to 4.50 and the standard deviation was between 0.41 to 1.26.

This indicated that the program was effective. When comparing each item, every item had its effectiveness as a high level. The average value of an organization's appropriateness in terms of screen, background and letter and convenience of use was 4.50; the highest and the standard deviation was 0.55. While the convenience of accessing and quitting the program, swiftness of the result processing, concordance between the program and its objectives and the convenience of installing the program had an average value of 4.33 and the standard deviations were 0.52, 0.52, 0.82 and 0.82 respectively. The average value of 4.17 and standard deviation of 0.75 are the benefit of a program in cognitive process diagnosis. The convenience of filling the information, appropriateness of data record and accuracy of the program processing had equal average value of 4.00 and the standard deviations were 0.89, 0.89 and 1.26 respectively. Finally, the lowest average value belonged to appropriateness of printing, result form and presentation of test results (3.83); the standard deviations were 0.41 and 0.75.

In Table 3, the results of evaluating the handbook of MCAT by the experts showed that program effectiveness was in the highest level; its average was 4.57; standard deviation, 0.50. The average value of each item was between 4.50 to 4.67, and the standard deviation was between 0.52 to 0.55. When comparing each item, the highest level belonged to an agreement between contents and processes of the program, appropriateness and clearness of the figures and effectiveness of the handbook; its average was 4.67 and standard deviation, 0.52. While the rest had an equal average value of 4.50 and the standard deviation was 0.55.

The results of trying to use the MCAT program for 2 times indicated that the buttons on the computer screen has some problems as follows;

The problems that the 3 math teachers and the 6 students from grade 7 faced were the font size and the screen was too small for data recording. Moreover, the teachers suggested that before using the program, returning button to the test main menu should be added. The researchers then improved everything the teachers and the students suggested.

There were some problems found in the second trial with 6 teachers and 18 students from grade 7. Teachers suggested that the test result button should not be placed at the bottom left because it is not appropriate with the eyes level; it should be left at the right top. The position of the ending testing was so close to the answer button; it should be removed to the bottom left. The item number should be removed from the bottom right to the top right so as to be proper with the eye level. There was only testing result report, but there was no item identification after testing; and the testing results of each issue should not be presented in the same screen. It should be separated. In the trial with students, no problem was found. The researchers had improved all points the teachers suggested

DISCUSSION

The results of the development of grade 7 students' algebra item bank by applying multidimensional item response theory model

MIRTM could be explained as follows. Items selection - when selecting the items into the bank, only a few items passed the criteria. That is items of Order and Graph passed 42% and Linear Equation with one variable passed 37%. It means that many items did not pass the criteria. This is because the Algebra was quite difficult and the item guessing value was also high. According to the research result of The Institute for Promotion of Teaching Science and Technology (Dechri and Kamparasiri, 2009), it was pointed out that the students had the lowest score in Algebra.

The results on developing of MCAT showed that Multidimensional Item Response Theory Model (MIRTM)

Table 2. The results of appropriation and accuracy of MCAT checked by the experts.

Evaluation criteria	$\overline{X}$	SD	Results
The concordance between the program and its objectives	4.33	0.82	High
The convenience of installing the program	4.33	0.82	High
An appropriate of organizing elements at the screen	4.50	0.55	High
An appropriate between background and letters	4.50	0.55	High
A convenience of accessing and quitting the program	4.33	0.52	High
A convenience of filling the information	4.00	0.89	High
A swiftness of the result processing	4.33	0.52	High
An accuracy of the program processing	4.00	1.26	High
An appropriation of presenting the testing results	3.83	0.75	High
An appropriation of a data record	4.00	0.89	High
An appropriation of printing and a result form	3.83	0.41	High
A convenience for applying	4.50	0.55	High
Benefits of a program in cognitive process diagnosis	4.17	0.75	High
Average	4.21	0.73	High

Table 3. The evaluation results of an accuracy and an appropriation of the MCAT handbook by the experts

Criteria	$\overline{X}$	SD	Results
An agreement between contents and the program's processes	4.67	0.52	Highest
An appropriation of Arranging contents	4.50	0.55	High
An appropriation of using the language	4.50	0.55	High
An appropriation of alphabets	4.50	0.55	High
An appropriation and clearness of the Figures	4.67	0.52	Highest
An appropriation of demonstrating the Figures	4.50	0.55	High
An effectiveness of the handbook	4.67	0.52	Highest
Total	4.57	0.50	Highest

was applied to develop Computerized Adaptive Testing for diagnosing grade 7 students' cognitive processes in learning Algebra.

The results from MCAT

The program that was designed for diagnosing testers' abilities at the same time including memorizing facts, understanding concepts, applying processes, analyzing concepts and entire ability, reporting the result immediately, saving time, supporting Largest Decrement in the Volume of the Bayesian Credibility Ellipsoid, and ending testing by a standard error of a low ability estimation or when 15 items are finished revealed that the time for testing was short. Testers might do 6-15 items owing to their ability. It could measure many of the testers' abilities at the same time and it contained an accuracy of estimating the testers' ability. Petersen et al. (2006) indicated that it was possible to estimate the 3 dimensions of a tester at the same time. Frey and Seitz (2009) also revealed that a computerized adaptive testing program by the item bank that was developed through Multidimensional Item Response Model was effective. It could decrease the items from Unidimensional Item Response by 30 to 50% and Classical Test Theory by 70% without decreasing accuracy.

The information function value of each item was high since it was from Largest Decrement in the Volume of the Bayesian Credibility Ellipsoid (Segall, 2010: 65-74). It leads to the highest posterior information matrix $\mathbf{I}_{i|s_{k-1}}$ by using 3 data sources: Inverse Prior Variance Covariance Matrix: $\boldsymbol{\varphi}^{-1}$, matrix aggregate $\mathbf{w}$ from the former item $\left(\mathbf{w}_{s_{k-1}}\right)$ and $\mathbf{w}$ matrix from the further item. The information function from the test result of this method depended on the relationship between a co-variance matrix and a number of the test dimension. If the relationship was complicated or there were a lot of dimensions, the information function would increase. This made the standard error to be less productive. The testers then

could finish the test without doing all items. From the development of Computerized Adaptive Testing, only 5 items of Linear Equation with one variable, and only 1 item of Order and Graph could obtain the ending criteria with standard error of a tester's ability estimation at lower than .30. It is concluded that the covariance matrix of Order and Graph cognitive process from this study was more complicated than the other one. Frey and Seitz (2009), Reckase (2009) and Diao and Reckase (2009) indicated that choosing the test item by Decrement in the Volume of the Bayesian Credibility Ellipsoid could decrease a large number of the items.

The criterion used in ending testing was ending it with standard error of a tester's ability estimation lower than .30 (Maneelek, 1997; SongSaeng, 2004; Gushta, 2003; Triantafillou et al., 2008). From applying CAT of this study, the testers could end testing with standard error of a tester's ability estimation lower than .30 with only 1 item of Order and Graph and 5 items of Linear Equation with one variable. This was not a suitable number of the test since the test items of CAT should be 25-36 items (SongSaeng, 2004; Gushta, 2003). This might affect the incredible of estimation. Therefore, the researchers adjusted the ending testing into 2 criteria: ending when the standard error of the testers' ability estimation of Order and Graph was 0.49 and Linear Equation with one variable was 0.13.

Studying the processing speed of a laptop computer with the processing unit (Pentium® Dual-Core CPU; band: T4400; processing speed: 2.20 GHz; memory: 2.0 GB; graphics card: Mobile Intel(R) 4 Series Express Chipset Family; memory of graphics card: 256 MB with resolution: 1280 × 720 pixels; and running on the operation system of Windows XP), it was found that the processing of choosing the next item of Linear Equation with one variable was slower than The Order and Graph since Largest Decrement in the Volume of the Bayesian Credibility Ellipsoid had many processes of an estimation which were from 3 sources of matrix, a complicated calculation. Moreover, more items indicated more slowness. In this case, the processing of Order and Graph was quicker than the other one by about 2 s since it had 59 items; while Linear Equation with one variable contained 104 items.

The evaluation results of MCAT's effectiveness and its handbook by the 6 experts revealed that the overall result in terms of application was in a high level since the program was developed by concepts, theories and various techniques. Only 1 item could direct more than 1 cognitive processes of a tester which could diagnose many skills of a tester; memorizing the fact, understanding the concept, applying the process and analyzing the concept. The program was suitable for use since it could report the result rapidly. Parshall et al. (2002) revealed that CAT could give a rapid response to a tester in giving scores and reporting the test results.

Moreover, the program gave accurate results. It was more suitable for a tester's ability to estimate and choose item. The mixed media also contributed in making the CAT more accurate. Parshall et al. (2002:23-25) also indicate that mixed media test could decrease the weakness of the tester, who has low reading ability. Questions related to expected question and the test process and result were significantly accurate.

From evaluating the CAT's handbook, the program could be applied at the highest level since the researchers tried it out and improved it. Therefore, it was effective.

Trying the program out in real situation

The researchers tried the program on teachers and students so as to find out whether the program was effective or not. Srisa-ard (1998) indicated that after developing the innovation, it should be tried personally on the samples before using it in real situation. This is done in order to find the mistakes in it and then to develop it. In this research, the program was tried out on 3 teachers and 6 students (personally), and then on a small group of 6 teachers and 18 students. Based on this, the researchers found that there were some problems about the font size, the position of the buttons, the unsuitability between the eye level and the buttons and the presentation pattern. These problems are related only to the satisfaction aspects of the program style, but there was no problem on the program's accuracy since the program was improved before the experts used it. In terms of the program's pattern, the researcher adjusted it in order to meet the user's satisfaction. This is related to the theory of the customer's satisfaction of Naumann and Giel (1995), which showed that customer's satisfaction depends on the quality of goods and services, price and overall image.

The results of 192 samples' opinions (18 teachers and 174 grade 7 students) on using MCAT indicated that the program's benefits, probability, suitability, accuracy and overall image were highly satisfactory. This was because the program was made based on principles, concepts, theories and techniques. Furthermore, it was proved by trying it out on the target groups two times. It was improved for good effectiveness before using it in a real situation. Therefore, it met the users' need.

Implementation of MCAT program by applying response item model developed in real situation

Teachers were able to use the program in teaching and learning by pre-test, while learning test and post-test. The developed program would be used to diagnose the individual student's ability. When students had taken the test, the program would be used to report the result of each student's ability followed by the cognitive intellectual process in 4 areas: 1) Fact remembering 2) Concept understanding 3) Applying approaches and 4) Concept analyzing. Teachers and students would immediately

receive the feedback as soon as they finish taking the test. After that, teachers would know the weaknesses of the students in every issue. Teachers could help students develop appropriately and purposefully by designing the tools such as practicing exercises, readymade lessons and so on; meanwhile, students would also know their weaknesses. They were able to improve themselves directly by revising and studying more contents or issues which affected their learning outcomes in the future relating to the result of research synthesis of Hattie (2009) on students' learning outcomes. It was found that the feedback teachers gave to the students highly influenced the result of students' learning outcomes (Buasuwan et al., 2011, Cited from Hattie, 2009).

Conclusion

(1) There were 163 items in Algebra item bank of grade 7 students including Order and Graph; 59 items have average value: -0.258; MDISC: 1.744 and MDIFF: 0.188; the average discrimination values in dimensions 1, 2, 3 and 4 were 1.137, 1.016, 0.568 and 0.931 respectively; 104 items of Linear Equation with one variable contain average value of d: -1.87; MDISC: 1.750; MDIFF: 1.058; the average discrimination values in dimensions 1, 2, 3 and 4 were 1.199, 0.718, 0.804 and 0.802 respectively.
(2) The MCAT program was effective on students. It could diagnosis the cognitive learning processes of grade 7 students namely recognizing, understanding, applying, analysis and overview. The program contained validity, reliability, benefit, probability, appropriation and accuracy which covered the diagnosis points of grade 7 students' abilities of learning algebra.

ACKNOWLEDGEMENT

This study was granted by The Promotion of Research in Higher Education Project, Office of the Higher Education Commission, Ministry of Education (2011-2012).

REFERENCES

Anderson LW, Krathwohl DR, Bloom BS (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational outcomes: Complete edition, New York: Longman.

Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR (1956). Taxonomy of educational objectives: Handbook on I: Cognitive Domian. New York: David MCkay.

Bock RD, Schilling SG (2003). IRT based item factor analysis. In M.du Toit (ed) IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Scientific Software International, Lincolnwood, IL, pp. 584–591.

Buasuwan P, Sudarats S, Thongthai W, Rungsayatorn C, Samahito C, Wichitputchraporn W (2011). Development of Thailand's Quality of Ecucation Framework through the Process of Critical Dialogue. J. Soc. Sci. Res., The Social Science Research Association of Thailand. pp.44-65.

Diao Q, Reckase M (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In: Weiss DJ (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. pp. 1-13 Retrieved on 8 February, 2011 from www.psych.umn.edu/psylabs/CATCentral/.

Frey A, Seitz NN (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. Stud. Educ. Evalu. 35:89-94.

Gushta MM (2003). Standard-setting Issues in Computerized-Adaptive Testing. Centre for Research in Applied Measurement and Evaluation Paper Prepared for Presentation at the Annual Conference of the Canadian Society for Studies in Education, Halifax, Nova Scotia, www2.education.ualberta.ca/educ/psych/crame/files/GushtaCSSE2003.pdf>. February 30th, 2011.

Haberman S (2008). When can subscores have value? J. Educ. Behav. Stat., 33(2):204–229.

Kanjanawasri S (2007). Modern Test Theories. 3rd ed. Bangkok: The Printing Press of Chulalongkorn University.

Ketterlin-Geller LR, Yovanoff P (2009). Diagnostic Assessments in Mathematics to Support Instructional Decision Making. Practical Assessment, Research Evaluation, 14(16):1-11. Available online: http://pareonline.net/pdf/v14n16.pdf.

Leighton JP, Gierl MJ (2007). Why cognitive diagnostic assessment? In Leighton, J. P., and Gierl. Cognitive diagnostic assessment. Cambrigde University Press: New York, pp. 205–237.

Maneelek R (1997). The Effect of Some Variables on Concurrent Validity and Item Number of Computerized Adaptive Testing. Dissertation's Thesis. Bangkok: Srinakharinwirot University.

McDonald RP (1999). Test Theory: A Unified Treatment. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Parshall CG, Davey T, Pashley PJ (2002). Innovative Item Types for Computerized Testing. In W. J. van der Linden & C.A.W. Glas (Eds.), Computerized Adaptive Testing: Theory and Practice (pp. 129-148). Netherlands: Kluwer.

Petersen MA, Groenvold M, Aaronson N, Fayers P, Sprangers M, Bjorne JB (2006). Multidimensional computerized adaptive testing of the EORTC QLQ-C30: Basic developments and evaluations. Quality of Life Research, 15:315–329.

Dechri P, Kamparasiri K (2009). Trends in International Mathematics Study 2007. Nontaburi: Sahamit Printing and Publishing.

Reckase MD, McKinley RL (1991). The Discriminating Power of Items that Measure More than One Dimension. Appl. Psychol. Measurement, 15(4):361-373.

Reckase MD, Martineau JA (2004). The Vertical Scaling of Science Achievement Tests. Paper Commissioned by the Committee on Test Design for K-12 Science Achievement Center for Education National Research Council. 15(4):1-25

Reckase MD (2009). Multidimensional item response theory. Springer Science+Business Media: New York.

Rupp AA, Templin J (2008a). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. Measurement, 6(4):219-262.

Rupp AA, Templin J (2008b). The effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. Educational and Psychological Measurement, 68(1):78-96.

Samejima F (1974). Normal Ogive Model on the Continuous Response Level in the Multidimensional Latent Space. Psychometrika, 39(1):111-121.

Segall DO (2010). Principles of Multidimensional Adaptive Testing. In Wim J. van der Linden and Cees AW Glas (Eds.). Elements of adaptive testing. Springer: New York Dordrecht Heidelberg London. pp. 57-75.

Sinharay S, Haberman S, Puhan G (2007). Subscores based on classical test theory: To report or not to report. Educational Measurement: Issues and Practice, 26(4):21-28.

Srisa-Ard (1998). Development of Teaching. 2nd Bangkok: Chomromdek.

SongSaeng K (2004). Test Information Functions in Computerized Adaptive Testing. Dissertation's Thesis. Bangkok: Srinakharinwirot University.

Triantafillou E, Georgiadou E, Economides AA (2008). The design and evaluation of a computerized adaptive test on mobile devices. Comput. Educ., 50:1319-1330.