

Evaluating Knowledge Structure-based Adaptive Testing Algorithms and System Development

Huey-Min Wu, Bor-Chen Kuo and Jinn-Min Yang

Research Center for Testing and Assessment, National Academy for Educational Research, New Taipei City, Taiwan
// Graduate Institute of Educational Measurement and Statistic, National Taichung University of Education,
Taichung, Taiwan // Department of Mathematics Education, National Taichung University of Education, Taichung,
Taiwan // lhswu@seed.net.tw // kbc@mail.ntcu.edu.tw // ygm@ms3.ntcu.edu.tw

(Submitted September 11, 2009; Revised January 26, 2011; Accepted March 31, 2011)

ABSTRACT

In recent years, many computerized test systems have been developed for diagnosing students' learning profiles. Nevertheless, it remains a challenging issue to find an adaptive testing algorithm to both shorten testing time and precisely diagnose the knowledge status of students. In order to find a suitable algorithm, four adaptive testing algorithms, based on ordering theory, item relational structure theory, Diagnosys, and domain experts, were evaluated based on the training sample size, prediction accuracy, and the use of test items by the simulation study with paper-based test data. Based on the results of simulation study, ordering theory has the best performance. An ordering-theory-based knowledge-structure-adaptive testing system was developed and evaluated. The results of this system showed that the two different interfaces, paper-based and computer-based, did not affect the examinees' performance. In addition, the effect of correct guessing was discussed, and two methods with adaptive testing algorithms were proposed to mitigate this effect. The experimental results showed that the proposed methods improve the effect of correct guessing.

Keywords

Adaptive test algorithm, Computerized adaptive test, Diagnostic test, Knowledge structure, Ordering theory

Introduction

During the last two decades, from the functional aspect, many computerized test systems have been developed for estimating abilities of examinees (Chang, Lin, & Lin, 2007; Guzman & Conejo, 2005; Lewis & Sheehan, 1990; Sands, Water, & McBride, 1997; Sheehan & Lewis, 1992; Wainer, 2000; van der Linden, 2000; Tao, Wu, & Chang, 2008; Yen, Ho, Chen, Chou, & Chen, 2010) or diagnosing students' learning profiles (Appleby, Samuels, & Treasure-Jones, 1997; Chang, Liu, & Chen, 1998; Hwang, Hsiao, & Tseng, 2003; Liu, 2005; Tsai & Chou, 2002; Tselios, Stoica, Maragoudakis, Avouris, & Komis, 2006; Vomlel, 2004; Yu & Yu, 2006). From the theoretical aspect, some of them are based on item-response theory (IRT) (Chang et al., 2007; Guzman & Conejo, 2005; Lewis & Sheehan, 1990; Sands et al., 1997; Sheehan & Lewis, 1992; Wainer, 2000; van der Linden, 2000; Yen, et al., 2010), some of them are based on artificial intelligence techniques such as Bayesian networks (Liu, 2005; Tselios et al., 2006; Vomlel, 2004), and others are based on knowledge structures. From the operational aspect, some of the computerized tests are adaptive and others are non-adaptive. The focus of this study is to construct computerized adaptive tests based on knowledge structures for diagnosing students' learning profiles.

The computerized adaptive test (CAT) can not only offer examinees customized items in accordance with their aptitudes or cognitive status, but can also shorten the test. The CAT based on IRT models can obtain efficient estimates of subjects' abilities, but it cannot provide the capability to diagnose subjects' cognitive concepts at a detailed level (Tatsuoka, Corter, & Tatsuoka, 2004; Yan, Almond, & Mislevy, 2004). Instead, knowledge structure- or artificial-intelligence-based adaptive tests can provide information about how well subjects performed on specific concepts, so they can achieve the diagnostic function (Appleby et al., 1997; Tatsuoka et al., 2004; Vomlel, 2004).

Diagnosys, developed by Appleby et al. (1997), is a knowledge-based-computer diagnostic test of basic mathematical concepts. In Diagnosys, a method was proposed to estimate the knowledge structure of examinees and then apply this structure to build the adaptive testing process. Chang et al. (1998) have proposed adaptive test algorithms to construct a computerized adaptive diagnostic test based on knowledge structures constructed by the domain experts. The results of these two papers exhibit that the proposed algorithms have the capability of decreasing the use of test items and are able to precisely diagnose the cognitive status of examinees. However, the impact of correct guessing on the diagnoses of concepts is not considered in these studies. Correct guessing means

that an item is answered correctly by guessing in multiple-choice tests. In knowledge-based adaptive tests, if an item is answered correctly by guessing, then all prerequisite items of it are assumed to have been answered correctly. But, in actual fact, these prerequisite items may not have been answered correctly. In that situation, the precision of diagnosing results would be decreased. Moreover, the impact of correct guessing in adaptive testing would be greater than that in non-adaptive testing such as the traditional paper-and-pencil test.

Tselios et al. (2006) used the Bayesian network to diagnose students' problem-solving strategies with two distinct problems. The results show that the Bayesian network can estimate students' problem-solving strategies very well, but it is not an adaptive test. Vomlel (2004) and Liu (2005) have proposed adaptive testing algorithms based on the Bayesian network. In their simulation study, the numbers of test items were 10 and 21, respectively. The experimental results show that the Bayesian network is a powerful tool to diagnose students' learning status; however, it is difficult and time consuming to find the optimal adaptive testing strategy when the test is long.

Ordering theory (OT; Airasian & Bart, 1973; Bart & Krus, 1973) and item relational structure theory (IRS; Takeya, 1991) were proposed for displaying the students' item structures. In previous studies (Bart & Krus, 1973; Takeya, 1991), OT and IRS were used for developing instruction sequences or learning progress indices. In this paper, OT and IRS are used to estimate knowledge structures of examinees and apply them to new adaptive test algorithms. One of the currently existing problems is that there are many knowledge-structure-based adaptive testing (KSAT) algorithms but no study to evaluate their performance. The performance of the adaptive testing algorithm, Diagnosys, domain experts, OT, and IRS is evaluated by using the simulation study; moreover, the effect of correct guessing in the multiple-choice tests are also explored in this study. In comparison to the algorithm proposed by Appleby et al. (1997) and the domain experts, our algorithms significantly reduce the length of time to take tests, and the algorithm with the best performance is selected to construct a computerized adaptive diagnostic test to be used in an actual Grade five diagnostic mathematics test. The experimental results show that the computerized adaptive diagnostic test has performed as expected.

Adaptive test algorithms based on knowledge structures

A hierarchy concept network, knowledge structure, introduced by Gagne (1977) as a way of defining prerequisite association of concepts, is the combination of named individual concepts, a specified level for each concept, and specified directed links between concepts that joins them together into a hierarchy. As shown in Figure 1, concept D is linked forwardly to concept C, which means that concept D must be mastered before concept C can be attempted; that is, concept D is a prerequisite for concept C.

By using this concept network, Appleby et al. (1997) proposed an inference mechanism (adaptive testing algorithm) that allowed the system to reduce the number of items that are administered in computerized adaptive diagnostic test. As shown in Figure 1, if the student gets concept D correct then it is inferred that he or she also knows its prerequisites (concepts F, G, H, and I). This algorithm in computerized adaptive diagnostic test can predict students' learning profiles by using fewer items than original paper-based tests.

The number of links has an impact on the use of test items. As the number of links increases, the use of test items decreases. In this paper, we propose adaptive testing algorithms with OT and IRS respectively.

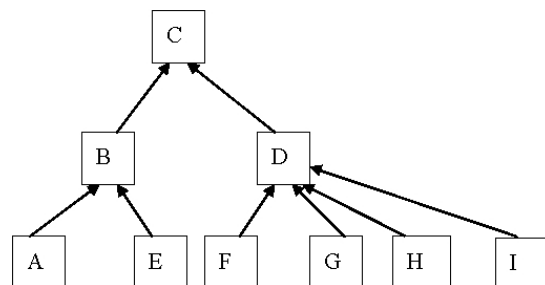


Figure 1. The knowledge structure

The domain experts' knowledge structure

Once a knowledge structure is constructed by practising teachers and domain experts, it is named as the domain experts' knowledge structure. The procedures for constructing a domain experts' knowledge structure are as follows. First, the domain experts define the important concepts of each unit by analyzing teaching materials and objectives. Second, after much discussion, the domain experts decide the sequence of the concepts development and relationships among these concepts to depict in a tree diagram the experts' knowledge structure for each unit. Figure 2 is an example of part of the domain experts' knowledge structure for a triangle unit of mathematics used in elementary schools of Taiwan. In the domain experts' knowledge structure, the upper-level concepts such as "find the isosceles triangle" are advanced concepts, while low-level concepts such as "find the right angle" are basic-level concepts. Generally, an item is developed to assess knowledge on a single concept. Diagnostic tests are developed by the concepts defined in the domain experts' knowledge structure.

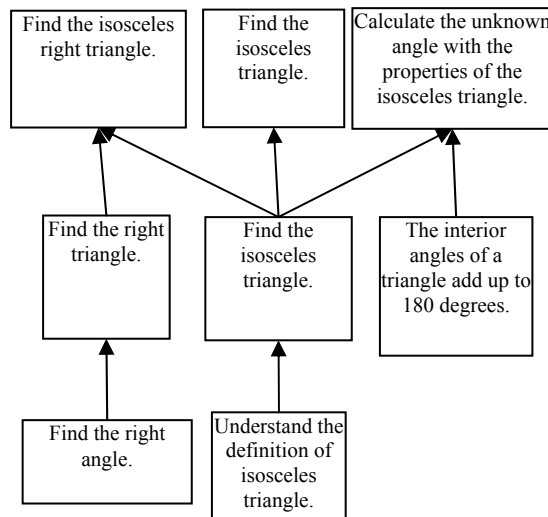


Figure 2. Part of the domain experts' knowledge structure for the triangle

Knowledge structure from Diagnosis

With Diagnosis, a paper-based pre-test is developed based on domain experts' structures and is then administered to collect responses from students. This data was applied to develop the inference mechanism as follows.

The relative frequencies of two concepts, A and B, are defined in Table 1. As shown in Table 1, f_{AB} represents the number of students with correct answers for both concept A and concept B. If $f_{AB} + f_{\overline{AB}} \gg f_{\overline{AB}} + f_{AB}$, then concepts A and B are equivalent and the relation is denoted as $A \leftrightarrow B$. Therefore, if students understand concept A, they will understand concept B as well, and vice versa. Moreover, if $f_{\overline{AB}} \gg f_{AB}$, then concept A could be linked forwards to concept B. The relation denoted as $A \rightarrow B$ means that A is a prerequisite to B. The important characteristic of the link $A \rightarrow B$ is twofold:

1. If the student gets an item on concept B correct, we can infer that she or he also understands concept A.
2. If the student gets an item on concept A incorrect, we can infer that she or he also does not understand concept B.

These two rules apply transitively across the structure according to the partial ordering given by the links. For example, for the network, $A \rightarrow B \rightarrow C$, if a student gets an item on concept C correct then we can infer that the student understands concept B due to direct inferences, but also A due to indirect transitive inferences. This algorithm allows the system to significantly reduce the number of items administered compared with a conventional test.

Table 1. Relative concepts frequency table

	A (correct)	A (incorrect)
B (correct)	f_{AB}	$f_{\bar{A}B}$
B (incorrect)	$f_{A\bar{B}}$	$f_{\bar{A}\bar{B}}$

Some inefficient problems are posed such as the definitions of ordering relation, equivalence and transition among concepts lack clarity while being inoperable. To improve these limitations, a threshold model is defined in this paper:

$$\text{If } \eta_{AB}^* = (f_{\bar{A}B} + f_{A\bar{B}}) / (f_{AB} + f_{\bar{A}\bar{B}}) < \varepsilon_{dia} \text{ then } A \leftrightarrow B.$$

$$\text{If } \lambda_{AB}^* = f_{\bar{A}B} / f_{A\bar{B}} < \varepsilon_{dia}, \text{ then } A \rightarrow B.$$

Ordering theory and item relational structure theory

In this paper, two other item ordering theories, OT and IRS are used for estimating knowledge structures of examinees and to develop new adaptive test algorithms. They are described briefly below:

Let $X = (X_1, X_2, \dots, X_n)$ denote a vector containing n binary item score variables. Each student taking an n -item test produces a vector $X = (x_1, x_2, \dots, x_n)$ containing 1 (correct) and 0 (incorrect). Then the joint and marginal probabilities of items on concepts A and are represented in Table 2.

Table 2 The joint and marginal probabilities of concepts A and B

		concept B		
		$X_B = 1$	$X_B = 0$	Total
concept A	$X_A = 1$	$P(X_A = 1, X_B = 1)$	$P(X_A = 1, X_B = 0)$	$P(X_A = 1)$
	$X_A = 0$	$P(X_A = 0, X_B = 1)$	$P(X_A = 0, X_B = 0)$	$P(X_A = 0)$
Total		$P(X_B = 1)$	$P(X_B = 0)$	1

For OT, let $\varepsilon_{AB}^* = P(X_A = 0, X_B = 1) < \varepsilon_{OT}$, usually $0.02 < \varepsilon_{OT} < 0.04$ (Airasian & Bart, 1973; Bart & Krus, 1973), concept A can be linked forward to concept B. The relation is denoted as $A \rightarrow B$ this means that A is a prerequisite to B. If $A \rightarrow B$ and $B \rightarrow A$, then the relation is denoted as $A \leftrightarrow B$ and it means concepts A and B are equivalent.

For IRS, Takeya (1991) proposed another index, r_{AB}^* , which is used to define the ordering relation from concept A to concept B. The definition of r_{AB}^* is

$$r_{AB}^* = 1 - (P(X_A = 0, X_B = 1) / P(X_A = 0)P(X_B = 1)) \geq \varepsilon_{IRS}$$

If $r_{AB}^* \geq \varepsilon_{IRS}$, then concept A can be linked forward to concept B. Usually the rule of thumb is to set $\varepsilon_{IRS} = 0.5$.

The performances of knowledge-structure-based adaptive testing (KSAT) algorithms

As mentioned above, four methods, Diagnosys, OT, IRS, and the domain experts, can be used to define knowledge structures. By applying these knowledge structures, the corresponding inference mechanisms (adaptive testing algorithm) are established. In this paper, we refer to them as knowledge-structure-based adaptive testing (KSAT). In this section, the performances of adaptive testing algorithms based on the four knowledge structures with different thresholds are compared and evaluated by using adaptive test simulation processes with a paper-based test dataset to determine the best algorithm. In these simulation processes, a paper-based test is taken owing to a limitation of computer equipment. The reason for using simulation is that there are hundreds of combinations of knowledge structure-based adaptive testing (KSAT) algorithms and thresholds. Finding a real computerized dataset for each combination is not feasible.

The use of test items and prediction accuracy of each combination are considered its performance. Mathematics definitions are defined in Table 3. As shown in Table 3, f_{ij}^{11} represents the frequency with which student j answered item i correctly, both in the simulated computerized adaptive diagnostic test and in the paper-based test; f_{ij}^{00} represents the frequency with which student j answered item i incorrectly, both in the simulated computerized adaptive diagnostic test and in the paper-based test. The prediction accuracy reflects a degree of similarity in the examinee's responses to the simulated computerized adaptive diagnostic test and the paper-based test. The use of test items is the average items administered to the examinees in the computerized adaptive diagnostic test. One of the goals of this paper is to find the best algorithm, which is able to achieve better prediction accuracy with fewer averaged use of test items. Once the best algorithm is determined by the training data, it is used in the actual computerized adaptive test.

Table 3. Definition of prediction accuracy and utilization of test items

		Simulated computerized adaptive diagnostic test	
		Correct (1)	Incorrect (0)
Paper-based test	Correct (1)	f_{ij}^{11}	f_{ij}^{10}
	Incorrect (0)	f_{ij}^{01}	f_{ij}^{00}

Prediction accuracy: $PA_{\varepsilon} = (1/Nn)(\sum_{j=1}^N \sum_{i=1}^n (f_{ij}^{11} + f_{ij}^{00}))$ where
 ε : threshold, $\varepsilon = 0, 0.01, \dots, 0.5$ for Diagnosys and OT; $\varepsilon = 0, 0.02, \dots, 1$ for IRS
 j : the examinee from test samples $j = 1, 2, \dots, N$
 i : the item $i = 1, 2, \dots, n$

Use of test items: $UI_{\varepsilon} = (1/N)(\sum_{j=1}^N n_j)$ where
 ε : threshold, $\varepsilon = 0, 0.01, \dots, 0.5$ for Diagnosys and OT; $\varepsilon = 0, 0.02, \dots, 1$ for IRS
 n_j : the number of items that are administered to the examinee j in the computerized adaptive diagnostic test
 j : the examinee from test samples $j = 1, 2, \dots, N$

To take an example from Figure 1, if a student has completed a paper-based test consisting of nine items, the response patterns are shown in Table 4. In the simulation KSAT process, if the student gets concept D correct then we can infer that the student also understood concepts F, G, H, and I, although they were not administered. Compared to the responses of the paper-based test, prediction accuracy and utilization of test items are calculated by the above-mentioned formula, $PA_{\varepsilon} = (8/9) = 0.89$, $UI_{\varepsilon} = 5$.

Table 4. Responses for a paper-based test and a simulated computerized adaptive diagnostic test

	Student responses								
	A	B	C	D	E	F	G	H	I
Paper-based test	0	0	0	1	1	1	0	1	1
Simulated computerized adaptive diagnostic test	0	0	0	1	1	<u>1</u>	<u>1</u>	<u>1</u>	<u>1</u>

0: incorrect 1: correct 1: inferred correct

Implementation of knowledge structure-based adaptive test (KSAT) system

The knowledge structure-based adaptive testing (KSAT) system has been implemented with PHP and MySQL on APACHE web servers. Figure 3 shows the architecture of the knowledge structure-based adaptive testing (KSAT), which consists of 10 modules: Account Management Module, Item Bank Management Module, Test Management Module, Competency Module, Diagnosis Module, Adaptive Item Selection Module, User-profile Database, Item Bank Database, Knowledge Structure Database, and Test Result Database.

The Account Management Module provides creation and management of user accounts. The functions of Item Bank Management Module include items or the knowledge structure of specific unit updates, modification, and management. The Test Management Module can set the approach of test administration. The Competency Module estimates the competency of individual students or groups. The Diagnosis Module diagnoses the knowledge states of the student by using the response pattern of the student. The Adaptive Item Selection Module can administrate tests according to different adaptive test algorithms. According to the experiment results, the knowledge structure estimated by the ordering theory has been used to construct the adaptive test algorithm that was placed in this module.

The following are several major interfaces of system.

The user management interface in Figure 4 is multi-functional. It allows new users to have access to creating new user accounts, creating multiple new user accounts, importing accounts from other sources such as Excel, and giving access to the database.

The test administration interface in Figure 5 displays the items and allows the examinees to answer the items presented. Since the KSAT system is an adaptive test, only one item per screen is presented.

The group profile interface in Figure 6 displays the group result of the exam. For example, in concept 5 of the interface, 13 students passed and 19 students failed test 1. Instructors can then take this information and understand the distribution of students' knowledge states and identify the strengths and weaknesses within a group. This information can be utilized for remedial instruction.

The individual profile interface is shown in Figure 7 and 8. Upon completion of the test, the student receives a personalized profile including name, scores, percentile, utilization of test items, date taken, and so forth. In Figure 14, the competency of the student for each concept in forms 1 to 3 is displayed.

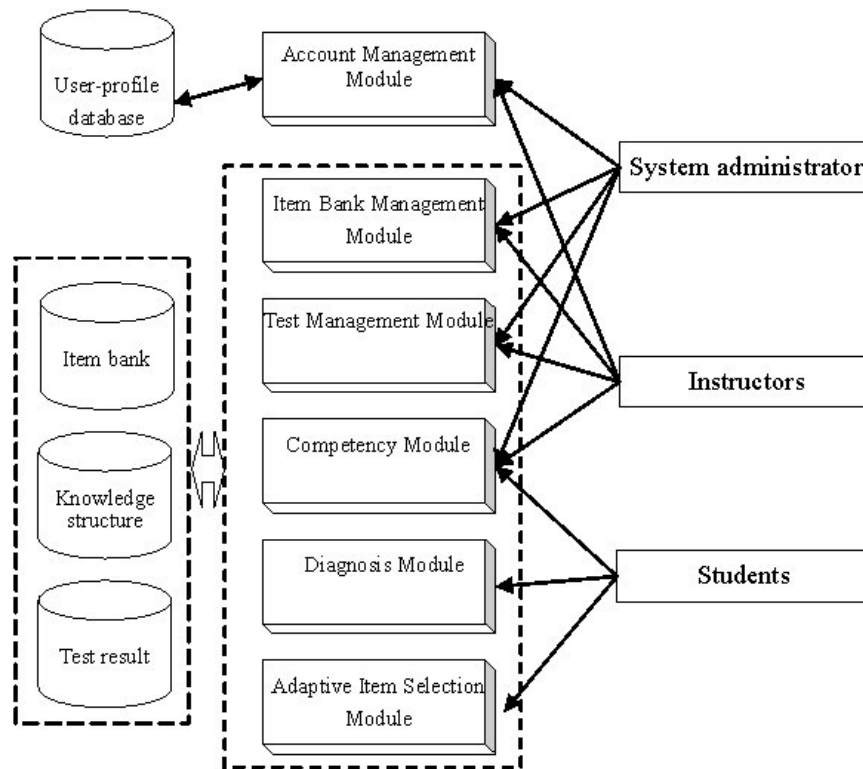


Figure 3. Architecture of knowledge structure-based adaptive testing (KSAT) system

User Management
[\[Create a new user account \]](#) [\[Create multiple new user accounts \]](#) [\[Import user accounts \]](#) [\[Access user database \]](#)

☆☆ Create a new user account ☆☆☆

Create a new user account

*ID:

*Password:

Re-type Password:

School location: County school Grade class

*Name:

*Gender: male female

I am a:

Birthday: 1950 - 01 - 01

An ID number:

Telephone:

Cell phone:

Address:

Email:


* required item

Figure 4. The user management interface

Version: A Volume: 9 Unit: 5 Title: triangle

Item :9

How many obtuse triangles are shown in the below figure?



4
 5
 6
 7

Figure 5. The test administration interface

The Group Profile
[\[Response patterns for the Class \]](#) [\[Statistics for the class \]](#)

☆☆ Response patterns for the Class ☆☆☆

Select class and test form:

Taichung heart 5 5 | Version: A Course: math Number:9 Unit:5 Title: Triangle

*Option1: show the statistics for test only
 show the statistics for unit

Option2: download .csv file

*required item

Taichung city, Heart elementary school grade 5 class 5 diagnostic profile [\[download .csv file\]](#)

concepts list	diagnostic profile	
	Form 1 pass-fail	Form 2 pass-fail
【concept1】 recognize the right angle	32-00	32-00
【concept2】 recognize the isosceles triangle	30-02	30-02
【concept3】 recognize the acute angle	14-18	29-03
【concept4】 recognize the obtuse angle	32-00	32-00
【concept5】 recognize the right triangle	13-19	27-05
【concept6】 recognize the isosceles triangle	32-00	32-00
【concept7】 the interior angles of a triangle add up to 180 degrees	30-02	30-02
【concept8】 recognize the acute angle	01-31	25-07
【concept9】 recognize the obtuse angle	32-00	32-00
【concept10】 recognize the isosceles righted triangle	17-15	16-16
【concept11】 find out the isosceles triangle	18-14	21-11
【concept12】 calculate the unknown angle with the properties of sosceles triangle	32-00	32-00
【concept13】 recognize the acute triangle	21-11	22-10
【concept14】 recognize the obtuse triangle	32-00	32-00
【concept15】 find out the acute triangle	22-10	25-07
【concept16】 find out the obtuse triangle	30-02	31-01
【concept17】 know the definition of congruent triangle	25-07	22-10

Figure 6. The interface for groups profiles

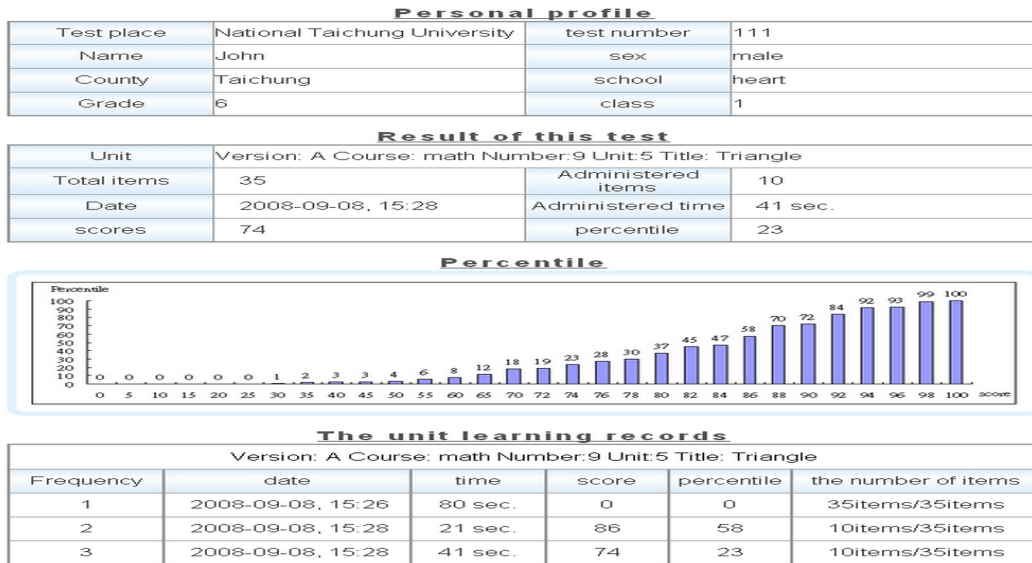


Figure 7. The profile for individual students interface

Concepts Diagnostic Profile

⊙ : pass the concept ×: fail the concept

Concepts List		Diagnostic Results		
		form1	form2	form3
【concept1】	recognize the right angle	× access	⊙	⊙
【concept2】	recognize the isosceles triangle	× access	⊙	⊙
【concept3】	recognize the acute angle	× access	× access	× access
【concept4】	recognize the obtuse angle	× access	⊙	⊙
【concept5】	recognize the right triangle	× access	× access	× access
【concept6】	recognize the isosceles triangle	× access	⊙	⊙
【concept7】	the interior angles of a triangle add up to 180 degrees	× access	⊙	⊙

Figure 8. The part of the diagnosis profile interface

Experiment 1 and results

The triangle unit of mathematics used in elementary schools of Taiwan was adopted to develop a paper-based test consisting of 35 items. The triangle mathematic test was administered to 660 selected fifth-grade students. As noted previously, four methods to define knowledge structures are mentioned. Three of the four methods require thresholds, \mathcal{E} , whereas the domain expert's structure does not require a threshold. The threshold effects of three algorithms (Diagnosis, OT, and IRS) on the prediction accuracy and use of test items were explored in this experiment. The responses of selected students were randomly divided into two parts, training samples and test samples. The training samples were applied to estimate the knowledge structures, and the test samples were used to estimate the prediction accuracy and use of test items. This process was repeated 50 times to obtain 50 sets of prediction accuracy and the use of test items. The averages of prediction accuracy and use of test items were used to represent the algorithm performance. The standard deviations of prediction accuracy and use of test items were used to evaluate the stability of the four algorithms. Training samples (TS) 10, 50, 100, and 200 were used to investigate the impact of the sample size on the prediction accuracy and on the use of test items of different algorithms.

Figures 9 to 18 present the prediction accuracy and the use of test items of different adaptive testing algorithms with different training sample (10, 50, 100, and 200). The scale of the horizontal axis of IRS (thresholds 0.02, 0.04, . . . 0.98) is different from those of Diagnosys and OT (thresholds 0.01, 0.02, . . . 0.50), so it is not displayed in the same graph. The horizontal axis represents the threshold, ε , and the vertical axis represents the prediction accuracy, PA_ε (Figures 9, 11, 13, 15, and 17) and the use of test items, UI_ε (Figures 10, 12, 14, 16 and 18). For example, in Figures 9 and 10, if the knowledge structure of Diagnosys with $\varepsilon = 0.08$ was applied, then $(PA_{0.08}, UI_{0.08}) = (0.821, 1.20)$, under the training samples, (TS) = 10. In Figures 17 and 18, if the knowledge structure of IRS with $\varepsilon = 0.58$ was applied, then $(PA_{0.58}, UI_{0.58}) = (0.896, 6.35), (0.952, 14.33), (0.970, 19.12), (0.984, 23.77)$ under the training samples (TS) = 10, 50, 100, and 200, respectively. The prediction accuracy and the use of test items of the algorithm based on domain experts' structure are 0.917 and 18, respectively. Since constructing the domain experts' structure does not need thresholds, it does not vary by thresholds.

Those figures show that:

1. Overall, the prediction accuracy and use of test items of Diagnosys and OT increase as the threshold decreases. The prediction accuracy and use of test items of IRS increase as the threshold increases.
2. Compared with the results from domain expert's structure $((PA, UI) = (0.917, 18))$, IRS $((PA_{0.32}, UI_{0.32}) = (0.923, 8.58))$, and OT $((PA_{0.08}, UI_{0.08}) = (0.922, 8.37))$ are able to achieve higher prediction accuracies with less use of test items.
3. For three test adaptive algorithms, Diagnosys, OT and IRS, the Diagnosys requires more training samples and higher use of test items to achieve the same or almost the same prediction accuracy in comparison to OT and IRS.
4. The performance of OT is less sensitive to the training sample size than that of IRS and Diagnosys.

For reducing the paper length without loss the generality, only three cases (case 1, case 2, and case 3) of means and standard deviations of prediction accuracies and their corresponding use of test items under the training sample size, 200 are displayed in Table 5. Case 1, case 2, and case 3 mean the prediction accuracy, 0.90, 0.92, and 0.94, respectively. The reason for choosing these cases in range of 0.90 to 0.94 is that this range is around the prediction accuracy of domain experts' structure and 0.94 is the maximum prediction accuracy that Diagnosys can achieve.

For example, in Diagnosys, when the average of prediction accuracy and its corresponding use of test items are 0.90 and 25.68, respectively, the standard deviations are 0.023 and 5.67, respectively. The range of standard deviations for prediction accuracy is 0.004 to 0.023, indicating that this simulation model is reliable. The lowest standard deviations of the prediction accuracy and the use of test items are all for the OT, so the OT has better performance on stability.

According to the results of the experiment, Diagnosys requires a large sample size and a larger use of test items to obtain better prediction accuracy, so it is not suggested for use. OT can obtain better prediction accuracy with less use of test items and training samples; hence OT is implemented into the KSAT system.

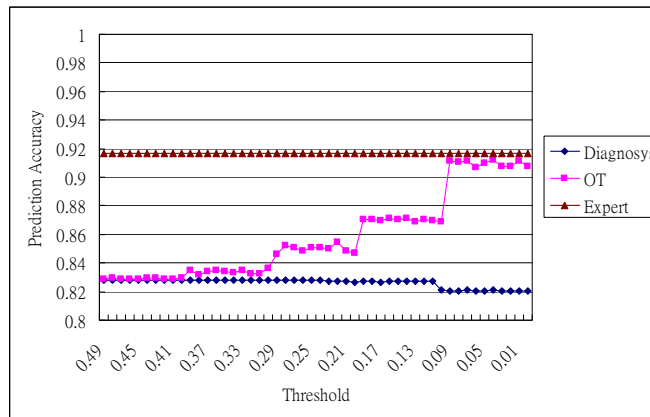


Figure 9. The prediction accuracy of Diagnosys, OT, and expert for training samples 10

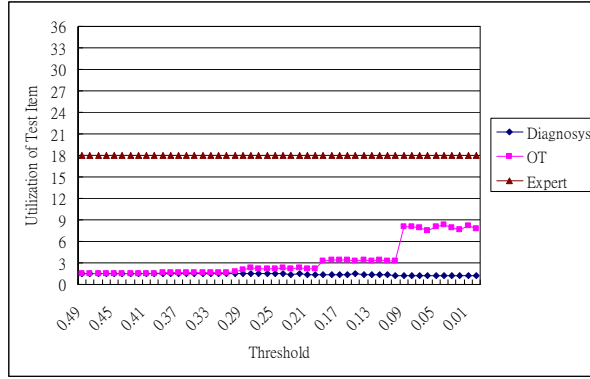


Figure 10. The use of test items of Diagnosis, OT, and expert for training samples 10

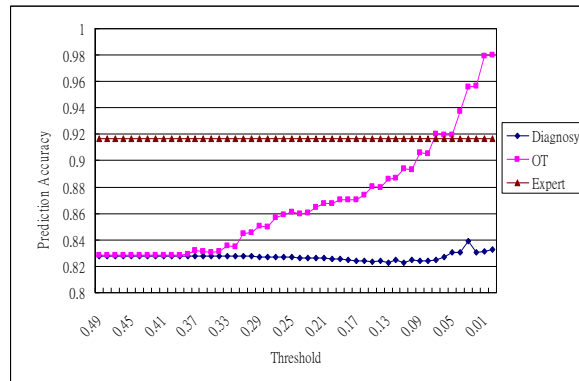


Figure 11. The prediction accuracy of Diagnosis, OT, and expert for training samples 50

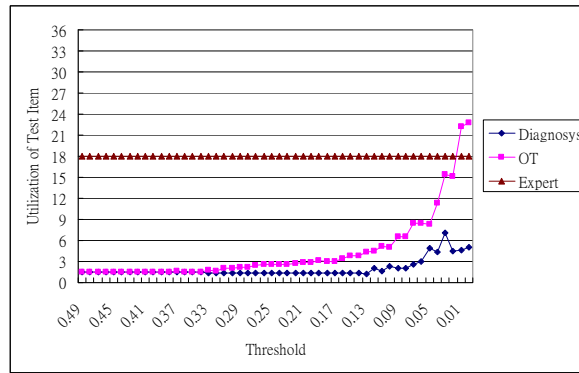


Figure 12. The use of test items of Diagnosis, OT, and expert for training samples 50

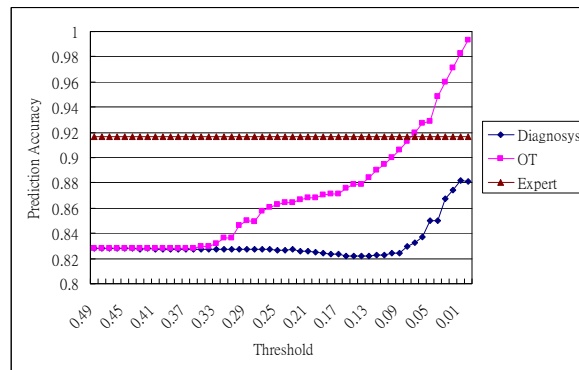


Figure 13. The prediction accuracy of Diagnosis, OT, and expert for training samples 100

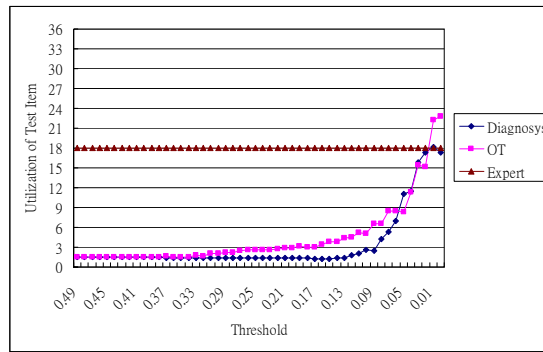


Figure 14. The use of test items of Diagnosys, OT, and expert for training samples 100

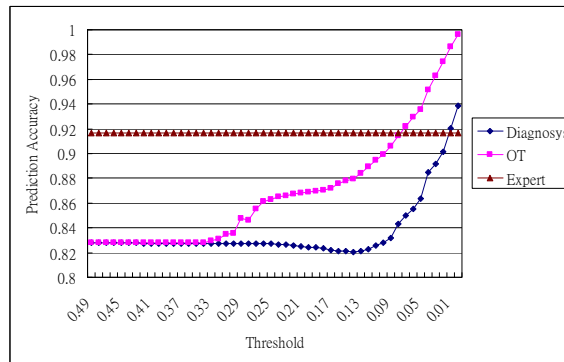


Figure 15. The prediction accuracy of Diagnosys, OT, and expert for training samples 200

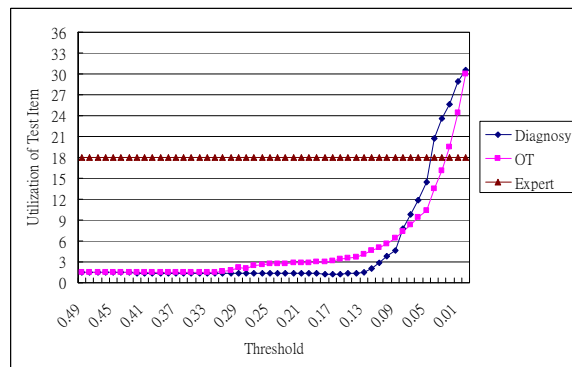


Figure 16. The use of test items of Diagnosys, OT, and expert for training samples 200

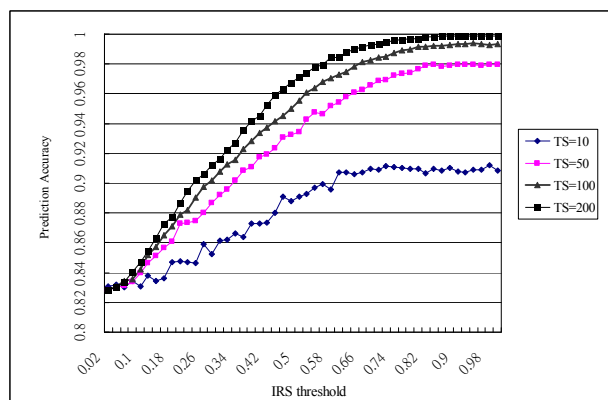


Figure 17. The prediction accuracy of the IRS for training samples 10, 50, 100, and 200

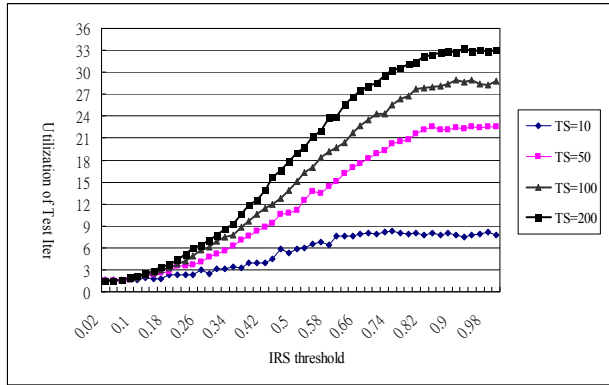


Figure 18. The use of test items of the IRS for training samples 10, 50, 100, and 200

Table 5. The means and standard deviations (in brackets) of the prediction accuracy and use of test items

		Case 1	Case 2	Case 3
Diagnosys	prediction accuracy	0.90 (0.023)	0.92 (0.015)	0.94 (0.016)
	use of test items	25.68 (5.67)	28.89 (2.43)	30.57 (2.17)
IRS	prediction accuracy	0.90 (0.010)	0.92 (0.007)	0.94 (0.007)
	use of test items	5.94 (1.003)	8.57 (1.163)	11.87 (1.400)
OT	prediction accuracy	0.90 (0.004)	0.92 (0.004)	0.94 (0.004)
	use of test items	5.60 (0.400)	8.37 (0.580)	10.39 (0.073)

Improvement of the correct guessing in KSAT algorithms

There are two major factors that affect the performances of adaptive testing algorithms. One is the theory to build knowledge structures and the other is the correct guessing of multiple-choice items. The effects of different theories are shown in the experiment 1, and we will discuss how to reduce the effect of the correct guessing in this section. In KSAT algorithms, if an item is answered correctly by guessing then all the prerequisite items of it will be assumed to be correct answers. This correct guessing would decrease the prediction accuracy of KSAT algorithms. Actually, the statistical nature of KSAT algorithms (especially OT) has the function to reduce the effect of correct guessing. Take OT as an example. OT, $\varepsilon_{AB}^* = P(X_A = 0, X_B = 1) < \varepsilon_{OT}$, if $P(X_A = 0, X_B = 1) \geq \varepsilon_{OT}$ (i.e., concept A is not a prerequisite of concept B), then the correct guessing only affects the prediction accuracy of concept B; otherwise the prediction accuracy of concept A will be influenced by the correct guessing. If the threshold is small, then the effect of the correct guessing decreases. But the use of test items will increase.

To improve this situation, two methods are proposed with KSAT. Take Figure 1 as an example, these two methods are described in the following.

Most difficult item (MDI) method: Suppose item C is answered correctly, then the most difficult item (suppose this is item B) in its prerequisite items will be presented to the examinee. If item B is answered correctly, then item C and its prerequisite items are recorded correct; otherwise, C is recorded and other prerequisite items should be taken by the examinee.

Prerequisite Item method (PI method): If item C is answered correctly, then the item with the largest number of prerequisite items (for example, item D) in C's prerequisite items will be presented to the examinee. If item D is answered correctly, then item C and its prerequisite items are recorded as correct; otherwise, C is recorded as incorrect and other prerequisite items should be taken by the examinee. If none of the prerequisite items of C have a prerequisite item, then a randomly selected item is applied to the examinee.

Experiment 2 and results

In this experiment, the performance of nine adaptive testing algorithms, Diagnosys, Diagnosys+MDI, Diagnosys+PI, OT, OT+MDI, OT+PI, IRS, IRS+MDI, and IRS+PI, were evaluated by using the same data set as Experiment 1. The use of test items and prediction accuracy were obtained by five-fold cross-validation. Results were presented in Table 6. For example, in Table 6, when the threshold was 0.01, the prediction accuracy of Diagnosys, Diagnosys+MDI, and Diagnosys+PI were 0.956, 0.996, and 0.992, respectively, and their corresponding use of test items were 32.64, 34.70, and 34.53. A Wilcoxon-Signed-Ranks test was used to compare the performances among nine models (Table 7). In Table 7, “Diag+MDI to Diag” indicates that the performance between original Diagnosys and Diagnosys with MDI was compared. Due to different thresholds, the performance of Diag+MDI, OT+MDI, and IRS+MDI were not explored. The results are as follows.

1. In Table 7, the results of the Wilcoxon-signed-ranks test revealed that Diagnosys, OT, and IRS, the prediction accuracies, adaptive testing algorithms with the most-difficult-item method (MDI) and prerequisite method (PI) both perform better than the original adaptive testing algorithms ($z = -3.422 \sim -3.409$, $p = 0.001$). Otherwise, in Diagnosys, Diagnosys+MDI outperform Diagnosys+PI ($z = -3.415$, $p = 0.001$); in OT, OT+MDI outperform OT+PI ($z = -3.066$, $p = 0.002$); in IRS, IRS+MDI outperformed IRS+PI ($z = -3.482$, $p = 0.000$). Overall, the performance of adaptive testing algorithms with the most difficult item (MDI) method was better than that of adaptive testing algorithms with the prerequisite method (PI method).
2. In Table 6, under the same (or almost the same) prediction accuracies, the use of test items in the proposed KSAT+MDI and KSAT+PI are fewer than those in the original KSAT algorithms. For example, in Table 6, when the prediction accuracies of Diagnosys, Diagnosys+MDI, and Diagnosys+PI are 0.945, 0.943, and 0.945, respectively and their corresponding use of test items are 31.85, 24.23, and 30.81. When the prediction accuracy of OT, OT+MDI, and OT+PI are 0.991, 0.991, and 0.991, respectively, their corresponding use of test items are 27.27, 26.25, and 26.18. When the prediction accuracy of IRS, IRS+MDI, and IRS+PI are 0.991, 0.991, and 0.991, respectively, their corresponding use of test items are 27.75, 26.38, and 27.47 (see grayed cells).
3. In Table 6, OT+MDI and OT+PI outperform Diagnosys+MDI, Diagnosys+PI, IRS+MDI, and IRS+PI at the same prediction accuracies. The only exception is at the prediction accuracy of 0.997, where the use of test items of OT+PI and IRS_MDI are 31.46 and 30.82, respectively (see bold cells).

Table 6. The prediction accuracy and use of test items (in brackets) of Diagnosys, OT, and IRS with MDI or PI

Diag threshold	Diag	Diag +MDI	Diag +PI	OT threshold	OT	OT +MDI	OT +PI	IRS threshold	IRS	IRS +MDI	IRS +PI
0.01	0.956 (32.64)	0.996 (34.70)	0.992 (34.53)	0.01	0.995 (29.98)	0.998 (31.51)	0.997 (31.46)	0.58	0.991 (27.75)	0.997 (30.82)	0.996 (30.68)
0.015	0.945 (31.85)	0.993 (34.50)	0.992 (34.45)	0.015	0.991 (27.27)	0.996 (29.18)	0.995 (29.16)	0.56	0.987 (26.44)	0.996 (30.16)	0.995 (30.03)
0.02	0.935 (31.02)	0.990 (34.25)	0.985 (34.05)	0.02	0.985 (24.41)	0.991 (26.25)	0.991 (26.18)	0.54	0.985 (25.20)	0.995 (28.89)	0.993 (28.73)
0.025	0.927 (30.22)	0.986 (33.92)	0.978 (33.58)	0.025	0.979 (22.14)	0.987 (24.07)	0.986 (23.94)	0.52	0.981 (23.50)	0.992 (27.59)	0.991 (27.47)
0.03	0.918 (29.31)	0.983 (33.50)	0.969 (32.93)	0.03	0.972 (19.42)	0.981 (21.59)	0.980 (21.42)	0.5	0.977 (21.82)	0.991 (26.38)	0.989 (26.21)
0.035	0.920 (29.50)	0.982 (33.43)	0.966 (32.62)	0.035	0.966 (17.06)	0.976 (19.13)	0.975 (19.01)	0.48	0.974 (20.79)	0.988 (25.17)	0.987 (25.04)
0.04	0.912 (28.14)	0.975 (32.43)	0.945 (30.81)	0.04	0.960 (15.59)	0.972 (17.68)	0.969 (17.43)	0.46	0.969 (19.47)	0.985 (23.85)	0.983 (23.73)
0.045	0.908 (27.20)	0.972 (31.83)	0.928 (29.31)	0.045	0.955 (14.50)	0.967 (16.54)	0.966 (16.32)	0.44	0.967 (18.52)	0.983 (22.86)	0.982 (22.72)
0.05	0.896 (23.55)	0.958 (28.52)	0.903 (24.97)	0.05	0.949 (13.30)	0.962 (15.43)	0.962 (15.27)	0.42	0.960 (16.47)	0.978 (20.62)	0.976 (20.41)
0.055	0.893 (22.40)	0.954 (27.36)	0.899 (23.76)	0.055	0.943 (11.95)	0.958 (14.23)	0.956 (14.06)	0.4	0.956 (15.17)	0.974 (19.29)	0.973 (19.21)
0.06	0.883 (19.27)	0.943 (24.23)	0.886 (20.39)	0.06	0.938 (11.07)	0.954 (13.31)	0.954 (13.29)	0.38	0.949 (13.54)	0.968 (17.38)	0.967 (17.20)

0.065	0.877 (17.56)	0.938 (22.55)	0.880 (18.71)	0.065	0.933 (9.99)	0.951 (12.45)	0.950 (12.36)	0.36	0.945 (12.66)	0.964 (16.42)	0.963 (16.23)
0.07	0.873 (16.37)	0.934 (21.40)	0.876 (17.52)	0.07	0.929 (9.36)	0.947 (11.79)	0.947 (11.70)	0.34	0.936 (10.99)	0.959 (14.52)	0.957 (14.34)
0.075	0.863 (13.57)	0.924 (18.63)	0.866 (14.73)	0.075	0.925 (8.73)	0.944 (11.10)	0.943 (11.03)	0.32	0.930 (9.97)	0.953 (13.34)	0.949 (13.09)
0.08	0.861 (12.70)	0.922 (17.76)	0.864 (13.85)	0.08	0.920 (8.18)	0.941 (10.67)	0.939 (10.46)	0.30	0.925 (9.04)	0.949 (12.21)	0.945 (11.92)

Note: Diag refers to Diagnosys.

Table 7. The results of Wilcoxon-Signed-Ranks tests among nigh models

	Diag +MDI v.s.Diag	Diag +PI v.s. Diag	OT +MDI to OT	OT+PI to OT	IRS +MDI to IRS	IRS +PI to IRS	Diag +MDI to Diag+PI	OT+MDI to OT+PI	IRS+MDI to IRS+PI
Z value	-3.422*	-3.422*	-3.409*	-3.410*	-3.411*	-3.409*	-3.415*	-3.066*	-3.482*

Note. * indicate the statistically significant at $\alpha = 0.01$

Evaluation of the KSAT system

To evaluate the performance of the KSAT system, an experiment has been conducted. This experiment aimed to evaluate the efficiencies of use of test items and prediction accuracy in administering the computerized adaptive test. One hundred and twenty-three students from fifth-grade classes of Taiwanese elementary schools participated in this experiment. The procedure was conducted as follows. First, all students received a knowledge-structure-based adaptive testing (KSAT) based on OT algorithm (threshold = 0.05). The content of the test was on the triangle unit, as mentioned above. Then, when the students completed the adaptive portion of the test, the system administered the rest of the 35 items in order to compute the prediction accuracy. Finally, the use of test items and prediction accuracy were computed.

After completion of the test, the results of the use of test items and prediction accuracy were 11.42% and 93%, respectively. The results show that the KSAT system can decrease the use of test items and are able to precisely diagnose the cognitive status of examinees.

These results are consistent with the results of OT case in the previous simulation experiment (the use of test items: 13; prediction accuracy: 95%). The two different interfaces, paper-based and computer-based, do not affect the examinees' performance in adaptive tests.

For exploring the performances of OT+MDI and OT+PI in this system, since the responses of all 35 items were available, this data set was applied to simulate OT+MDI and OT+PI processes. This simulated result shows that the use of test items and prediction accuracy of OT+MDI and OT+PI processes were (11.3, 94%) and (11.2, 94%), respectively. This implies OT+MDI and OT+PI have better performance than original OT, which is similar to the results of experiment 2.

Discussions and conclusions

In this paper, some traditional item ordering theories (OT and IRS) that were used to develop instruction sequences or learning progress indices were applied to develop the computerized adaptive testing processes. The performances of the adaptive testing algorithms based on the item structures constructed by OT, IRS, Diagnosys, and domain experts were evaluated. Three findings were found from the experimental results. First, OT and IRS based KSAT algorithms provide better prediction accuracy with less the use of test items. Second, OT-based KSAT algorithm is less sensitive to the training sample size. Third, the estimation error of OT method is less than others and this means that the diagnostic results estimated by OT-based KSAT is more stable. From the theoretic view of OT,

$\varepsilon_{AB}^* = P(X_A = 0, X_B = 1)$ is the probability of violating the ordering relationship of $A \rightarrow B$; that is directly related to the prediction error. From the definition and explanation of IRS in Takeya (1991), r_{AB}^* is designed to be a coefficient that has the benefits of both ε_{AB}^* and the correlation coefficient of items A and B. However, this modification reduces the direct relationship with the prediction error and affects the performance of IRS. The definition of Diagnosys, $\lambda_{AB}^* = f_{\overline{AB}} / f_{AB}$ shows that the error frequency $f_{\overline{AB}}$ is divided by f_{AB} , and this cause the relationship between λ_{AB}^* and the prediction error is reduced. From these observations, it is reasonable that OT has the best performance.

The performance of knowledge structure-based adaptive testing (KSAT) is affected by the correct guessing. Two methods, most difficult item method (MDI) and prerequisite item (PI) method, were proposed to deduce the possibilities of guessing. The experimental results show that both methods can improve the effect of correct guessing and have better performances than original methods.

Since OT has the best performance, it was selected to implement the KSAT system. The performance of the KSAT system shows that under the 93 percent prediction accuracy, the use of test items is 11.42. That is, on average, students need only complete one third of 35 items in the original paper-based exam when the KSAT system is applied. This result is close to that of the simulation study in experiment 1 and shows that the simulation process adopted in this study is valid and suitable.

From the above discussions, this study has three contributions. First, some evaluation methods for KSAT algorithms were applied to find the best adaptive testing algorithm among domain experts, OT, IRS, and Diagnosys, and OT-based KSAT algorithm has the best and stable performance. Second, two methods, the most difficult item method (MDI) and the prerequisite item (PI) method were proposed to improve the effect of correct guessing in KSAT algorithms. Finally, an OT-based adaptive testing system was developed and evaluated. Upon completion of the adaptive test, a diagnosis profile about the student's state of learning or understanding was provided to do the subsequent actions, such as tailored instruction or remediation in applied educational settings. Another two directions are considered in the future study. First, OT, IRS, and Diagnosys were used to analyze the ordering relationship of dichotomous items, those methods for polytomous items will be considered in the next step. The second is to develop constructed response items and their automatic scoring mechanism to enhance the function of the KSAT system.

Acknowledgement

The authors would like to thank the National Science Council of the Republic of China and the National Taichung University of Education for financially supporting this research under contract numbers: NSC-92-2521-S-142-003, NSC 97-2511-S-142-004, NSC 98-2410-H-142-005-MY2, NSC-100-2410-H-656-007 and 98T202-3.

References

- Airasian, P.W., & Bart, W.M. (1973). Ordering theory: A new and useful measurement model. *Journal of Educational Technology*, 5, 56–60.
- Appleby, J., Samuels, P., & Treasure-Jones, T. (1997). Diagnosys—A knowledge-based diagnostic test of basic mathematical skills. *Computers Education*, 28(2), 113–131.
- Bart, W.M., & Krus, D.J. (1973). An ordering theoretic method to determine hierarchies among items. *Educational and Psychological Measurement*, 33, 291–300.
- Chang, S.H., Lin P.C., & Lin Z.C. (2007). Measures of partial knowledge and unexpected responses in multiple-choice tests. *Educational Technology & Society*, 10(4), 95–109.
- Chang, K.E., Liu, S.H., & Chen, S.W. (1998). A testing system for diagnosing misconceptions in DC electric circuits. *Computers and Education*, 31, 195–210.
- Gagne, R.M. (1977). *The Conditions of Learning*. Holt, Rinehart & Winston, New York, 3rd ed.
- Guzman, E., & Conejo, R. (2005). Self-assessment in a feasible, adaptive web-based testing system. *IEEE Transactions on Education*, 48(4), 688–695.

- Hwang, G.J., Hsiao, C.L., & Tseng, C.R. (2003). A computer-assisted approach to diagnosing student learning problem in science course. *Journal of Information Science and Engineering*, 19(2), 229–248.
- Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computer mastery test. *Applied Psychological Measurement*, 14, 367–386.
- Liu, C.L. (2005). Using mutual information for adaptive item comparison and student assessment. *Educational Technology & Society*, 8(4), 100–119.
- Sands, W.A., Water, B.K., & McBride, J.R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.
- Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement*, 16, 65–76.
- Takeya, M. (1991). *New item structure theorem*. Tokyo: Waseda University.
- Tao, Y.H., Wu, Y.L., & Chang, H.Y. (2008). A practical computer-adaptive testing model for small-scale scenarios. *Educational Technology & Society*, 11(3), 259–274.
- Tatsuoka, K. K., Corter, J. E. & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, 41(4), 901–926.
- Tsai, C.C., & Chou, C. (2002). Diagnosing students' alternative conception in science. *Journal of Computer Assisted Learning*, 18, 157–165.
- Tselios, N., Stoica, A., Maragoudakis, M., Avouris, N., & Komis, V. (2006). Enhancing user support in open problem solving environments through Bayesian network inference techniques. *Educational Technology & Society*, 9(4), 150–165.
- van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computer-adaptive testing: Theory and practice* (pp. 27–52). Boston: Kluwer.
- Vomlel, J. (2004). Building adaptive tests using Bayesian networks. *Kybernetika*, 40(3), 333–348.
- Wainer, H. (2000). *Computerized adaptive testing: a primer* (2nd ed). NJ: Lawrence Erlbaum Associates.
- Yan, D., Almond, R. G., & Mislevy, R. J. (2004). *Comparisons of cognitive diagnostic models* (ETS Research Report RR-04-02). Princeton, NJ: Educational Testing Service.
- Yen, Y.C., Ho, R.G., Chen, L.J., Chou, K.Y., & Chen, Y.L. (2010). Development and evaluation of a confidence-weighting computerized adaptive testing. *Educational Technology & Society*, 13(3), 163–176.
- Yu, S.C., & Yu, M.N. (2006). The relationships among indices of diagnostic assessment of knowledge Structure and S-P chart analysis. *Journal of Education & Psychology*, 29(1), 183–208.