DESIGNING A VIRTUAL ITEM BANK BASED ON THE TECHNIQUES OF IMAGE PROCESSING

Wen-Wei Liao, Rong-Guey Ho
Graduate Institute of Information and Computer Education
National Taiwan Normal University
Taipei, Taiwan
abard@ice.ntnu.edu.tw, hrg@ntnu.edu.tw

ABSTRACT
One of the major weaknesses of the item exposure rates of figural items in Intelligence Quotient (IQ) tests lies in its inaccuracies. In this study, a new approach is proposed and a useful test tool known as the Virtual Item Bank (VIB) is introduced. The VIB combine Automatic Item Generation theory and image processing theory with the concepts of figural tests and Computerized Adaptive Testing (CAT). It is believed that this tool will assist in improving traditional figural tests – in terms of solving previous issues relating to item exposure and allowing a figural test to be more easily developed.

INTRODUCTION

With the development of technologies, the computer has evolved into a tool that can improve the accuracy and efficiency of tests. In effect, computers have largely transformed the way in which testing has been conducted over the years. Computer-Based Testing (CBT) has been adopted both in Taiwan and overseas. Examples of CBT include the Graduate Record Examination (GRE), the Graduate Management Admissions Test (GMAT), Test of English as a Foreign Language (TOEFL), and On-line Computer Basic Competence Test of High School and Vocational School Students (http://www.onlinetest.org/).

In comparison to the CBT, the Computerized Adaptive Testing (CAT) is a more complex form of testing. A CAT system chooses items for a given examinee based upon the examinee's responses to earlier items, as well as estimating one's ability according to his/her responses. As a result of the reduction in both testing time and testing items, many studies have since focused on the application of CAT (Ho, 2000).

Both CAT and CBT have security problems. Theoretically, these problems can be solved only when the item bank is so huge to the extent of infinity. As a matter of fact, creating a huge item bank is a lot of work as it needs a lot of manpower, resources, time and budget (Yu, 1993). Therefore, it has been an important task for researchers to look for different ways to solve test security problems.

Today, two major indicators for test security are the item overlap rate and the item exposure rate (Tsai & Kuo, 2005). The item exposure rate refers to the frequency at which a certain item appears for all examinees, while the item overlap rate refers to the frequency at which the same item appears for two examinees. As higher item exposure rate and item overlap rate mean higher risks for test security, it is necessary to put the two indicators under control in a real test. However, today's research only focuses on the control of the item exposure rate, while there has not been an effective way to control the item overlap rate. Accord to studies, the item exposure rate and the item overlap rate are not independent but interdependent (Tsai & Chen, 2005).

To solve item exposure and overlap problems, Virtual Item Bank (VIB) is proposed in this study. VIB does not attempt to control the item exposure rate and the item overlap rate, but it replace item for "object" and "rules" in the VIB. The VIB combines "object" and "rules" into items required in the test. These items not only satisfy the needs of CAT and CBT, but also solve the test security problem by minimizing the risks of item exposure and overlap problems.

This study will plan to use VIB in figural tests. Figural tests are comprehensive mental ability testing tools for children and the illiterate. However, it is acknowledged that building a figural test can be rather challenging (Cronbach, 1990). It analyzed items in APMs for their problem-solving rules using this technique to create a VIB. This item bank can solve problems concerning the item exposure rate and the item overlap rate. It can also help test editors to solve the test security problems and accurately calculate examinees' ability in a more efficient and safer way.

Test Security

CBT or CAT is administrated by selecting items from item bank. However, test security problems concerning item overexposure will arise when a great number of examinees have participated in the test over time. We can assess the test security by two key indicators: one is the item exposure rate and the other is the item overlap rate.

Copyright © The Turkish Online Journal of Educational Technology

Initially, we protected test security by randomly selecting items for a more even item distribution (Chang, 2003). However, no desirable results were seen in this method. Therefore, some researchers solely focused on the control of the item exposure rate in the hopes that this problem could be solved. Most discussed control method in their study is SH Procedure (Sympson & Hetter procedure) proposed by Sympson & Hetter (1985). This method was done by using the ability distribution of a group of simulated examinees to control the item exposure rate prior to the test. To achieve better control, the ability distribution of this group of examinees should be similar to that of the actual test takers. To make this happen, different exposure control parameters were used in examinees with different levels of ability.

Chang (2003) proposed SHC (Sympson & Hetter conditional procedure). SHC is a kind of control mode which divides examinees with different levels of ability into different groups, obtains the exposure control parameters of each item in different levels of ability, and combines the parameters into an exposure control matrix as the basis of exposure control in a real test. For fewer examinees with higher and lower ability, the maximum expected exposure parameter should be adjusted higher; On the contrary, for more examinees with medium ability, the maximum expected exposure parameter should be adjusted lower to increase the usage rate of the item (Chen, 2007). Other methods which can control item exposure rate includes Stocking & Lewis (1995) unconditional multinomial (SL) procedure , Stocking and Lewis (1998) conditional multinomial (SLC) procedure, Davey & Parshall procedure (DP, 1995), SH online procedure with freeze control (SHOF) (Chen, 2005). However, these methods do not take item overlap rate into consideration so that item overlap problems remain.

According to studies, item exposure rate and item overlap rate are not independent but interdependent (Chen, 2004). That was when Chen& Lei (2005) developed SHT that controlled both the item exposure rate and the item overlap rate to complement SH. Like SH, SHT requires pre-simulated exposure parameters as they both have time-consuming and test scenario problems. To solve this problem, Chen, Lei& Liao (2008) extended SHT into SHTO so that the efficiency of controlling item exposure problems can be dramatically enhanced by controlling item exposure rate and item overlap instantly online without having to pre-simulate exposure parameters. Nevertheless, either SHT or SHTO can only control the item overlap rate between two examinees. In fact, an examinee can obtain test information from more than one person. Therefore, it is necessary to control the item overlap rate between one prospect examinee and a group of examinees who have already taken the test. To broadly control item overlap rate, Chen (2008) proposed SHGT control method. Similar to SHTO, SHGT can instantly control item exposure rate and overlap rate on line. They differ from each other in that SHTO can only control item overlap rate between two examinees, while SHGT can do so for one prospect examinee and α past examinees ($\alpha \geq 1$).

Although researchers have come up with different ways to control both item exposure rate and item overlap rate, test disclosure remains a problem when there are too many users over time (Chang, 2003). Thus, some researchers use Automatic Item Generation (AIG) technique to generate items. AIG has not been used until recently (Irvine & Kyllonen, 2002) although it has been proposed for 30 years. There are numerous approaches for generating items using a computer (Millman & Westman, 1989), but they generally require the availability of an item model. An item model (Bejar, 2002; Drasgow et al., 2006) is a general prototypical representation of the items to be generated. Furthermore, each component of an item model can contain both fixed and variable elements (Lai, Alves & Gierl, 2009). Using item model, AIG can generate countless items to solve item exposure rate and overlap rate problems. However, this method cannot be applied in CBT or CAT as it cannot accurately calculate examinee's ability.

Designing CAT and CBT is challenging as it takes a lot time and resources to create the item bank. According to a study conducted by Chen (2007), only 78 research papers done by PhDs and graduate school students in Taiwan are on tests (10 on traditional Computer-based Testing, CBT; 35 on Computerized Adaptive Testing, CAT; 33 on Online Testing). It is even rare to see papers on figural testing. Therefore, it is an important job for researchers to help test editors to design the item bank for figural tests using fewer manpower and resources in a shorter time. This study will develop a new technique, called VIB, based on AIG and using item exposure rate and item overlap rate as indicators. This study will use this technique in figural tests to solve both item exposure and overlap problems. To generate desirable distracters in figural tests, this study combines Content-Based Image Retrieval technology to generate options with higher distractibility.

Content-Based Image Retrieval

Generally speaking, figural tests were more difficult for test editors to generate than text tests. In the selection verification, examinees paid full attention to the accuracy of the selection and the problem introduced by the option. As multimedia technology advances, this study would use content-based image retrieval to help

Copyright © The Turkish Online Journal of Educational Technology

examinees solve the problem of selection verification. Image comparison has been applied in many fields such as identity authentication, surveillance, human-computer interface, multimedia etc. In this research, content-based image retrieval techniques in image processing would be employed. Also, the main parts of the figure would be identified in order to perform data mining. The concepts and methods of content-based image retrieval are described below:

(1) Formula without considering color characteristics.
The characteristic vector is used in the computation to represent the figure, as shown below:

fi=(i1,i2,i3,... in) ……..……..…….(i)

f is the characteristic vector of the figure, and n is the code for the content characteristic. The similarities of two figures are obtained by computing the Euclidean distance of the characteristic vector (as shown in Formula i). The smaller the value, the more similar the two figures and vice versa.

$d(Q,I)=\sqrt{\sum_{j=1}^{n}(f_j^Q - f_j^I)^2}$ ……..……..……..……(ii)

While d(Q ,I) is Euclidean distance of the characteristic vector of figure I and Q (Berretti, Bimbo & Pala, 2000; Euripides, Petrakis & Evangelos, 1999).

(2) Formula that considers color characteristic:
Mehtre, Kankanhalli & Lee (1998) proposed a solution to consider the figure color and shape together to calculate the figure similarity. The methods are described as follows.

Step 1: Find the color clusters in the figure. The formula for the color distance is shown in Formula iii. While clustering 400 x 400 figure color, the minimum threshold of the color distance between each cluster was set to 50.

Color distance=$\sqrt{(\Delta R)^2+(\Delta G)^2+(\Delta B)^2}$ ……... (iii)

Step 2: Find the clusters in the figure. In step 1, we categorize color clusters into layers. In step 2, we mark the shape cluster of each layer, and line up the shape cluster according to pixels in each layer pattern. If the number of pixels in the shape cluster is less than 50, then this shape cluster is omitted. In order to avoid mistaking thin lines for clusters, the minimum density (see Formula iv) of shaper cluster as the shape threshold is set.
density,

$$\rho=\frac{population\ \ of\ \ \ Cluster}{(l_{\max})^2} \quad \text{……..……(iv)}$$

$l_{\max}$ =max(||x2- x1||, ||y2- y1||) (x2, y1) and (x2, y2) are corner points of shape cluster.

(3) Similarity calculation:
Using the formula for the color and shape distance (Formula iv and v), the similarity of the color and shape can be calculated. Next, use Formula vi to compute the similarity of the two features (Finlayson, Chatterjee & Funt, 1996).

$coldis(C_i^Q, C_j^I)=\sqrt{(R_i^Q-R_j^I)^2+(G_i^Q-G_j^I)^2+(B_i^Q-B_j^I)^2}$ ……..……(v)

Figure Q has m color cluster and p shape cluster. Figure I has n color cluster and q shape cluster.

$$shpdis(C_i^Q, C_j^I)=\sqrt{\sum_{i=1}^{7}(m_i^Q-m_i^I)^2} \quad \text{……..……(vi)}$$

i is moment invariant.

$$D(Q,I)=\omega_1\psi_1+\omega_2\psi_2+\omega_3\psi_3+\omega_4\psi_4 \quad \text{……..……..……..(vii)}$$

$$\psi_1=\sum_{i=1}^{\max(m,n)} cdist(C_{c,i}^Q, C_{c,Pc(i)}^I), \quad \psi_2=\sum_{i=1}^{\max(m,n)}\sqrt{(\lambda_{c,i}^Q-\lambda_{c,Pc(i)}^I)^2}$$

$$\psi_3=\sum_{i=1}^{\max(p,q)} shpdist(C_{cs,i}^Q, C_{cs,Ps(i)}^I), \quad \psi_4=\sum_{i=1}^{\max(p,q)}\sqrt{(\lambda_{s,i}^Q-\lambda_{s,Ps(i)}^I)^2}$$

Copyright © The Turkish Online Journal of Educational Technology

$\omega_1, \omega_2, \omega_3, \omega_4$ are weighted index.

Pc is the closest color cluster assignment function, and can map every color cluster i of image Q to the closest color cluster Pc(i) of image I. Formulas that consider the color instead of color characteristics helped generate suitable answers in this study. We proposed the process of building VIB along with the above studies. This process does not only apply to figural tests, but to all types of tests. This technique and process will also provide best practices for researchers working on testing theories to solve test security problems.

METHODS

This study has developed research tools and VIB based on the principles of test design and APM materials. Research tools are used to generate VIB by defining object and its composition method. The development procedure of this study is shown below:

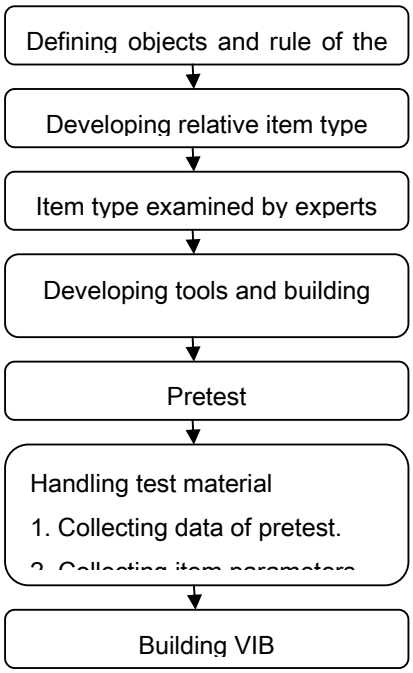


Figure 1, The development procedure of this study.

Participants

This study was conducted for a total of 310 six graders in 10 classes of an elementary school in New Taipei City, Taiwan. These participants must have basic computer literacy and have been involved in APM testing. This study will test these students using our self-developed CAT system and VIB. Test results will be analyzed with APM scores to demonstrate the feasibility of VIB.

Defining Testing Object and Rules

The word "object" defined in this study means the fundamental element of forming a test. In a Cube Counting Test, the "object" is the "Cube". In Four Arithmetic Operations, the "object" is numbers. In Test of Nonverbal Intelligence (TONI) (Brown, Sherbenouv & Johnsen, 1982), some basic shapes in their original form such as Circle, Triangle, Square, Rectangle, Parallelogram, Trapezium, Ellipse and Sector are the "objects" in a TONI test. Rules refer to how an object works. In a Cube Counting Test, piling the boxes is the "rule" of the test. In Four Arithmetic Operations, adding, subtracting, multiplying, and dividing are the "rule". In a TONI test, the variation of the shapes is the "rule".

This study uses APM as the material to build VIB. There are 36 formal items in APM suitable for examinees of 12 years old or above with higher intelligence. As this is not an adaptive test, an item contains a 3×3 matrix stem and several distracters. There are known shapes in the first eight boxes of the graphic matrix, while there are not any shapes in the narrow box in the lower right. Examinees must carefully observe the difference and variation of the shapes in the boxes in the horizontal or vertical direction, find out the correlation among shapes and their variation rules, and decide which shape to fill in the blank box according to their correlation or rules. This test contains a series of analytical and reasoning items in which graphic matrix will progressively change their directions horizontally or vertically. During the process, changes may involve increase and decrease of the shape

Copyright © The Turkish Online Journal of Educational Technology

size, addition or subtraction of elements, flip-over, turn-around or progressive changes in other forms.

In short, this is correlation search and target management (Carpenter, Just, & Shell, 1990) and the rule defined in this test. APM puts a great emphasis on test-retest reliability (the reliability every other four week is and .71~.78 and the split-half reliability is .59~70). In terms of validity, the correlation between APM and the graphic IQ test is .51~.75; the correlation between APM and math scores in junior high schools is .45~.72. APM is good for both individual and group tests and also an ideal test tool to analyze human fluid intelligence (Yu and Huang, 1990). In this study, the "objects" in APM are known shapes, such as triangle, circle, square, and graphic elements randomly generated by the computer. As for the "rule", there had been a lot of research conducted on APM. Arendasy and Sommer (2005) concluded six "rules" as shown below using geometry as the element:

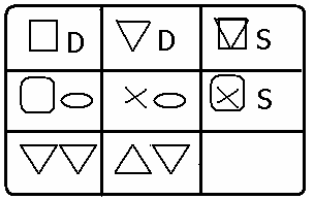
(1) Addition: Add the same graphic elements in the first two boxes and put them together in the last box.
(2) Intersection: Only the graphic elements in the first two boxes in the same position will be reserved for appearance in the last box.
(3) Seriality: Move progressively in a fixed direction (clockwise or counter-clockwise) in the box.
(4) Completeness: The same shape elements of all types appear in all boxes at the same frequency. For example, there are three shapes that need to appear three times.
(5) Neighborhood: Shape elements in adjacent boxes will appear in a connected position.
(6) Subtraction: The same shape elements in the first box should appear individually in the second and third box. Their appearance should not repeat or skip.

Freund, Hofer & Holling (2008) proposed five principles for figural matrix in APMs:
(1) Complete Addition: Piling elements stay in their original position, while other shape elements are all added and combined together in the last box.
(2) Addition-1 Element: Shape elements that appear only once in the previous box will be reserved in the last box. Elements that have already appeared more than twice will not be shown.
(3) Addition-2 Elements: Shape elements that only appear twice in the previous box will be reserved in the last box. Elements that do not appear twice will not be shown.
(4) Progression-Position: Shape elements move their position in a fixed direction, such as clockwise or counter-clockwise.
(5) Progression-Form: For example, each shape element will appear twice in all narrow boxes. Therefore shape elements that only appear once will be shown in the last box.

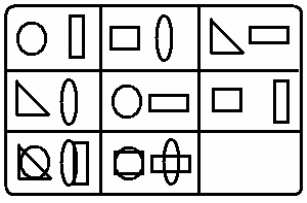
This study combines the rules proposed by the above two scholars and uses image processing technique (And, Or, Xor, Sub) to create the rule for virtual item banks. For example, "adding" is when two "objects" are processed by Or; "subtracting" is when two "objects" are processed Xor. This study uses image processing technique to execute the "rule" in the above APM test and save them into VIB.

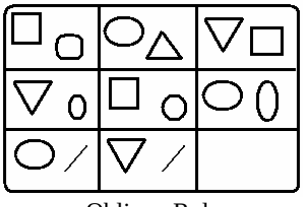
Developing Relative Types

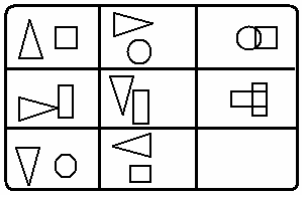
This study uses APM as the material through self-observation to define 12 rules as shown below. These 12 rules can create 48 different rules when working with four operations (And, Or, Xor, and Sub) of image processing technologies.

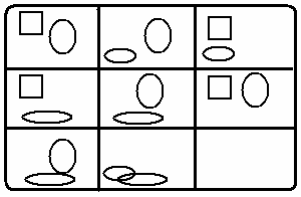


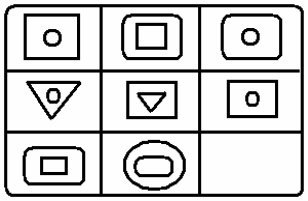
Addition Rule


Diagonal Rule


Oblique Rule


Allocation Change Rule


Quantitative Pair Wise Progression Rule


Size Change Rule

Copyright © The Turkish Online Journal of Educational Technology

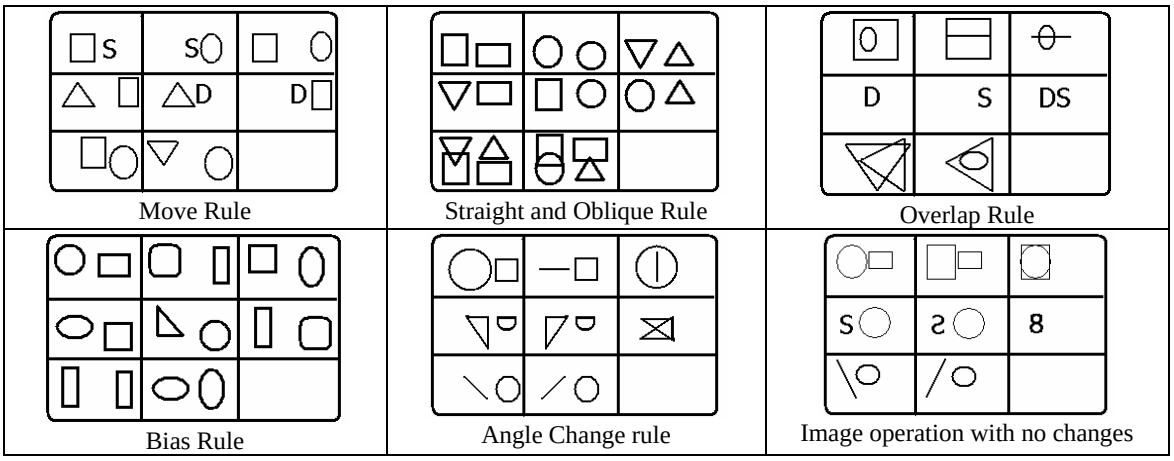


Figure 2, Item generation rules without image processing operation.

Building an Item Bank for pretest

When building a pretest item bank, the bank should contain requires four materials as shown below:

(1) Examinee's personal information
(2) Test paper which includes test number, test name, and correct answers
(3) Examinee's question-answering response form for recording the examinee's test name and question question-answering response.
(4) Correlation Form of examinee's question-answering response versus correct answers to check if the examinee answer questions correctly so that the examinee can give an appropriate feedback.

We asked experts and scholars to validate if the 48 item types conform to the content indicator in relative space ability testing and check if the meaning of the questions is clear enough. Following the validation, appropriate corrections will be made. A total of 48 questions will be designed based on the 48 question types and they are numbered 1 to 48.

Pretest

Pretest taker: This study focuses on users who can use the computer to browse webpage and use the mouse to click on the question options. Basically, the age of examinees is not limited. However, if the examinee is too young to use the computer or have difficulty reading question sentences, assistance will be needed to help him answer the questions. This study involved 207 six-graders of an elementary school in New Taipei City, Taiwan, in the pretest. Individual test item and the whole test system were corrected or adjusted based on the student's test results. The table shown below is the descriptive statistics of the difficulty level and discrimination index of the pretest. The average difficulty level is the average discrimination index is .68, which is of medium levels

Table 1, Descriptive Statistics of Difficulty Level and Discrimination Index of Pretest

	N	Minimum	Maximum	Mean	Std. Deviation
Test Difficulty Level	207	.3100	.6811	.5101	.0687
Test Discrimination Index	207	.4121	.9012	.6801	.0872

Developing research tools and building VIB

This study has developed VIB and the research tools specifically for them. Their functions are described as below:

(1) Virtual Item Bank (VIB)
Virtual item banks will replace traditional item banks and become the item source for CAT or CBT. There are only "objects" and "rules" in virtual item banks. The parameters of each item are obtained from the pretest.

(2) Research tools
This study has developed four systems: 1.item rule definition subsystem, 2. item generation subsystem, 3. answer retrieval subsystem and 4.CAT system. Each subsystem has different tasks and functions and is described below:

I. Item rule definition subsystem:

Copyright © The Turkish Online Journal of Educational Technology

This system helps test editors to define “objects” and “rules”. Test editors can define “objects” and the “rule”. From the system screen, we can see the position of each “object” can be defined when a figural test is being designed in the system. We can also select the method in which how these objects are formed. The item parameters generated by the “objects” and “method” will be obtained in the pretest. “Objects” and “rules” defined by this system will be recorded in VIB. Items will be generated by the VIB along with the Item Generation Subsystem. The functions of the Item Generation Subsystem are described as below.

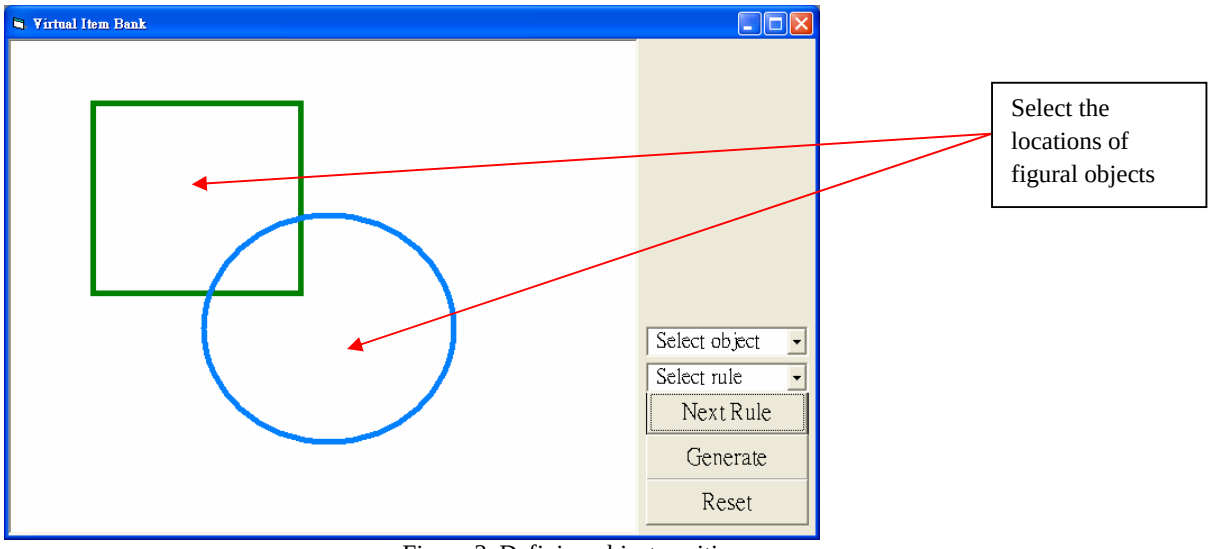


Figure 3, Defining object position

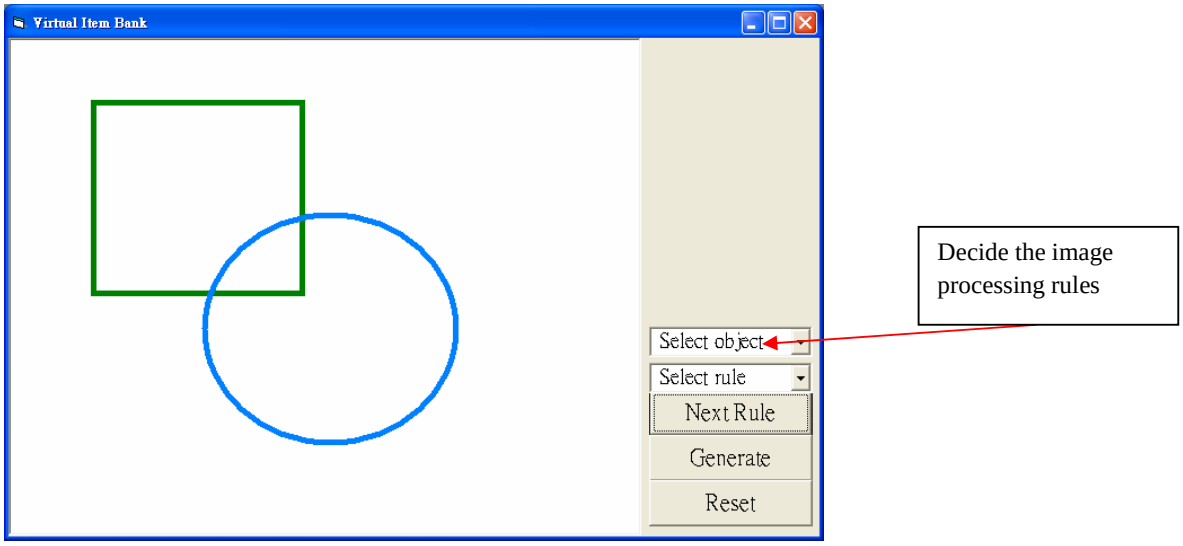


Figure 4, Defining the composition “rule” of the objects

Copyright © The Turkish Online Journal of Educational Technology

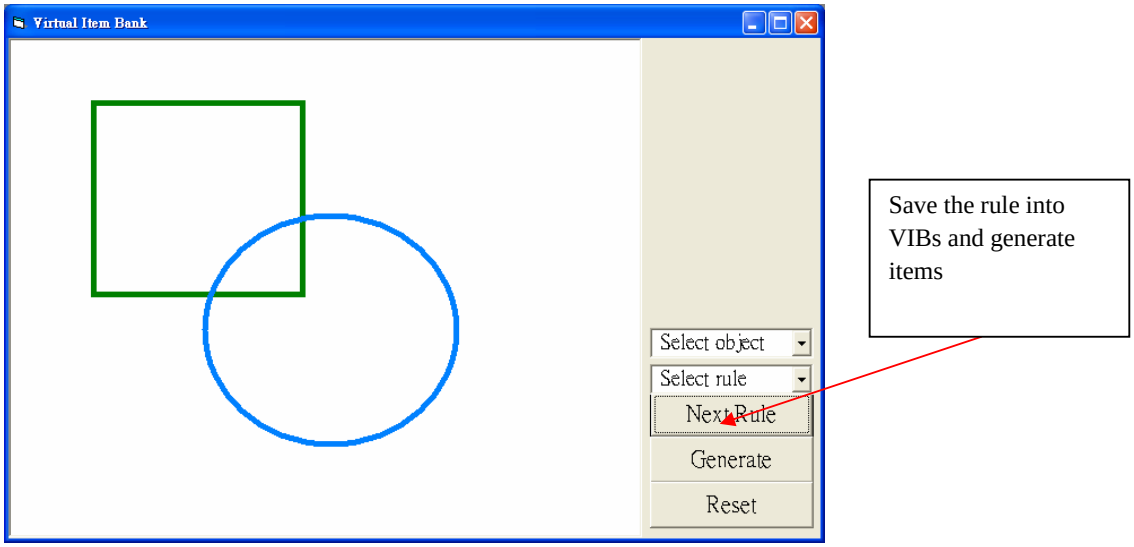


Figure 5, Save "objects" and "rules" into VIB and get ready to input the next item

II. Item generation subsystem:

The main function of this system is to read the "objects" and "rules" in VIB and generate relative items. The system can transform an object, for example, from "circle" to "hexagon" or from "square to "triangle". Theoretically speaking, with random transformation of objects and formation of "rules", there should be countless items in VIB. However, not every object transformation is reasonable. Therefore, there are three additional functions in this system that are responsible for forming and validating "objects" and "rules" to make sure all items generated by the system are reasonable and effective.

Function 1: Image processing function

This function can combine objects through image processing technique according to the "rules" defined by test editors. Besides, this function can make subtle changes to the size and position of each "object" to make the best combination for an effective item.

Function 2: Data retrieve function

This function can take items through the content-based image retrieval process. It can avoid multiple answers to one question or similar questions that may confuse examinees. When validating items, the function uses the above-mentioned content-based image retrieval technique to analyze the similarity between two items. Items with higher similarity are not to be used to avoid risks of item exposure.

Function 3: 3*3 Matrix Control function

This function focuses on the control of the 3*3 Matrix of item stems. It generates items by making changes to the "rules" defined by test editors from left to right, "up" to "down", or "diagonally".

III. Options retrieval subsystem:

Alternative options of each item were generated by image comparison. First, we computed the RGB value of the figures' pixel as the characteristic value. Then, we saved the figure characteristic into a 2-dimension matrix, and compared it with figures in the database. The similarities of the two figures were used to calculate the Euclidean distance (as shown in formula viii) of the characteristic value, and we select the lowest three as the alternative option.

$$d(Q,I)=\sqrt{\sum(f^{Q}-f^{I})^{2}} \quad \text{.....(viii)}$$

IV. CAT System:

The main function of CAT system is to select appropriate items for examinees and evaluate examinees' ability based on IRT model. In producing items, CAT system is only an application interface, and does not perform image process, item design or retrieval. These tasks are done by VIBS, and the results are sent back to CAT system to administer tests. In terms of ability evaluation, this system uses IRT to process. The psychometric model includes Rasch model, Two-parameter

Copyright © The Turkish Online Journal of Educational Technology

models (2PL model), and Three-parameter models (3PL model). The formulas are below:

Rasch Model $$P(\theta) = \frac{1}{1+e^{-1.7(\theta-b)}}$$

2PL model $$P(\theta) = \frac{1}{1+e^{-1.7a(\theta-b)}}$$

3PL model $$P(\theta) = c + \frac{1-c}{1+e^{-1.7a(\theta-b)}}$$

Among them, θ represents the examinees' ability; $p(\theta)$ represents the chances of examinees with θ ability answer an item correctly; b is the difficulty parameter; a is called the discrimination parameter which allowing an item to discriminate differently among the examinees; and c, the guessing parameter, represented the probability that an extremely low ability examinee would get the item correct. Since the system simplifies factors that affect the items, the Rasch model is used in this study. The functions of the research tools are described as follows.

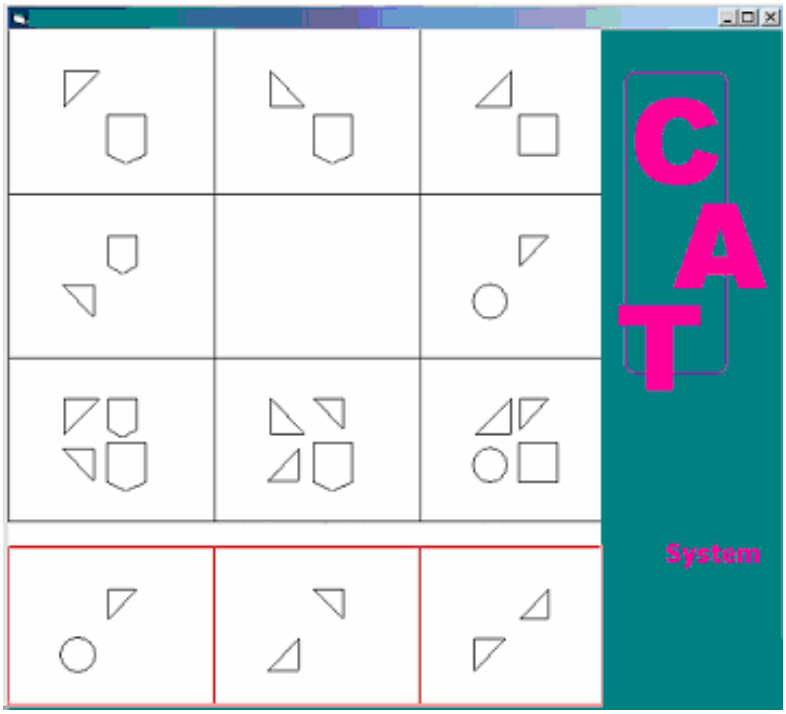


Figure 6, CAT System

RESULTS

The result was discussed in three parts: first was the descriptive statistics; second was item exposure and overlap rate validation and finally was VIB validation. This is to show that the feasibility of the VIB can be validated and security problems of the item bank can be addressed.

Descriptive Statistics

In this study, a VIB has been built specifically for figural testing. This VIB contains 48 rules. Each rule is composed of an image processing operation and a problem-solving rule. The primary parameter of this study was item difficulty parameter. An experiment deploying an online CBT system was designed to collect and estimate the items created by the VIB. 207 elementary school students participated in this experiment. The results of this experiment are presented in table 2. The structure of the CBT system and the results of the estimation are discussed as follow. An online CBT system is use to collect data of the item-generation rules in the VIB. An instruction example of this system would show to examinees before testing start. When examinees finished the test, the results would be transfer to the server and be analyzed in a short time.

Copyright © The Turkish Online Journal of Educational Technology

Table 2, Difficulty parameter estimations of item generation rules

D1	Difficulty	D2	Difficulty	D3	Difficulty	D4	Difficulty
1	0.21	13	0.46	25	0.21	37	0.65
2	0.56	14	0.44	26	0.39	38	0.4
3	0.49	15	0.49	27	0.4	39	0.63
4	0.08	16	0.71	28	0.48	40	0.35
5	0.8	17	0.27	29	0.22	41	0.34
6	0.12	18	0.22	30	0.83	42	0.43
7	0.4	19	0.19	31	0.29	43	0.42
8	0.39	20	0.31	32	0.72	44	0.65
9	0.4	21	0.83	33	0.7	45	0.61
10	0.27	22	0.27	34	0.6	46	0.6
11	0.34	23	0.76	35	0.63	47	0.6
12	0.4	24	0.45	36	0.6	48	0.51

D1 : Item generation rule are composed of SUB operation and 12 processes
D2 : Item generation rule contains OR operation and 12 processes
D3 : Item generation rule are consisted of AND operation and 12 processes
D4 : Item generation rule are composed of XOR operation and 12 processes

The item difficulty parameters created by the same rule were closed to each other. The results of the experiments are described as table 3.

Table 3, Results of item difficulty parameters generated by the rule A, B, C

Rule A	Difficulty	Rule B	Difficulty	Rule C	Difficulty
A-1	0.63	B-1	0.5	C-1	0.28
A-2	0.69	B-2	0.5	C-2	0.39
A-3	0.67	B-3	0.56	C-3	0.41
A-4	0.7	B-4	0.53	C-4	0.38
A-5	0.73	B-5	0.59	C-5	0.34
A-6	0.72	B-6	0.52	C-6	0.16
A-7	0.72	B-7	0.52	C-7	0.25
A-8	0.64	B-8	0.59	C-8	0.42
A-9	0.81	B-9	0.59	C-9	0.34
A-10	0.72	B-10	0.48	C-10	0.38

Table 4, Standard deviation of difficulty parameter of rule A, rule B, and rule C

Rule A	Rule B	Rule C
0.051218	0.04158	0.082226

The VIB generated items with similar item difficulty parameters by the same rules. The result indicated that the item difficulty parameters created by the same rule were closed to each other, which meant that the VIB is a powerful tool, and it can solve the problem of item exposure.

The Item Overlap Simulation

In this study, an item overlap simulation was conducted. According to the item overlap rate (given in formula ix), when max length of the test = 48, subjects = 30000, the simulation results are as follows.

Copyright © The Turkish Online Journal of Educational Technology

$$R_t = \frac{T_O / C_2^N}{(\sum_{i=1}^{N} L_i) / N} = \frac{2T_O}{(N-1)\sum_{i=1}^{N} L_i} \quad \text{.....(ix)}$$

R_t – test overlap percentage

T_O – the total numbers of items that both subjects overlap

L_i – the test length of the ith subject

Table 5, Results of the item overlap rate simulation

Item overlap rate (R)	$2.43488E^{-10}$
Mean of test length	36.5078
Mean of Theta-Estimated	-0.106
Mean of SE	0.4023

Table 6, Use frequency (times) of each item-generation rules

D1	frequency	D2	frequency	D3	frequency	D4	Frequency
1	22965	13	25020	25	21256	37	22966
2	25055	14	22870	26	22550	38	19741
3	24651	15	23302	27	20359	39	22747
4	26672	16	21622	28	24849	40	23579
5	21806	17	22000	29	20948	41	22762
6	23813	18	22271	30	20147	42	21263
7	21389	19	24464	31	24956	43	21142
8	25197	20	23895	32	21689	44	21860
9	22124	21	20891	33	22063	45	23353
10	22814	22	20581	34	24960	46	23111
11	23651	23	21064	35	23918	47	21411
12	21251	24	24848	36	24838	48	24549

Table 7, Item exposure rate of each rule

D1	Item of Exposure Rate	D2	Item of Exposure Rate	D3	Item of Exposure Rate	D4	Item of Exposure Rate
1	0	13	0	25	0	37	0
2	0	14	0	26	0	38	$1.0128e^{-4}$
3	0	15	0	27	0	39	0
4	$3.74925E^{-05}$	16	0	28	0	40	0
5	0	17	0	29	0	41	0
6	$4.19939E^{-05}$	18	0	30	0	42	0
7	0	19	0	31	0	43	0
8	0	20	0	32	0	44	0
9	0	21	0	33	0	45	0
10	0	22	0	34	0	46	0
11	0	23	0	35	0	47	0
12	0	24	0	36	0	48	0

The simulation results proved that VIBS solves the problems of item exposure.

Copyright © The Turkish Online Journal of Educational Technology

Validating VIB

After the VIB was built, a test was administrated for 310 examinees. We used examinees' APM scores as the external criterion and the total score of computerized figural testing in this study for Pearson Product-Moment Correlation analysis. The examinees' scores in "computerized figural testing" and the descriptive statistics of their APM scores are shown in Table 8. The two scores are positive correlated to a desirable level (r = 0.683, n = 310, p = .000). It means the result in computerized figural testing is relevant to the examinees' IQ scores calculated by APM.

Table 8, Descriptive Statistics of "Computerized Figural Testing" Score and APM's Score

	Mean	Std. Deviation	N
Computerized Figural Testing score	38.68	5.450	310
APM' Score	29.40	3.620	310

Table 9, Correlated Coefficient of "Computerized Figural Testing" Score and APM's Score

		Computerized Figural Test	APM Score
Computerized Figural Test	Pearson Correlation	1	.683**
	Sig. (2-tailed)		.000
	N	310	310
APM Score	Pearson Correlation	.683**	1
	Sig. (2-tailed)	.000	
	N	310	310

**. Correlation is significant at the 0.01 level (2-tailed).

CONCLUSION

In this study, we proposed a new technique called, "VIB", to address test security problems. This technique integrates AIG, Content-base Image Data Retrieval, item exposure rate control, and item overlap rate control to do so. Using VIB to administrate a test can rule out item exposure and overlap problems. Using VIB can also precisely calculate examinee's real ability without an error.

To validate the study, we conducted a test using APM as the material to build a VIB for figural testing and using CAT system to link the VIB. This study found out the combination rule of APM tests some research on APM and uses image processing operations, such as And, Or, Xor, and Sub to establish these rules. In this study, using image processing techniques helped us to easily and quickly generate items.

To address the technical problems on distracters, the purpose of this study aims to prevent similar distracters that may confuse examinees. We used the content-based image retrieval technique to analyze the similarity of two options. Options with higher level of similarity will be removed by VIB. Likely, items with similar stems will also be taken out by VIB so that the items will make more sense to examinees.

Working with all the above techniques, we developed research tools that included item rule definition subsystem, item generation subsystem, answer retrieval subsystem and CAT system. Using these tools, test editors can easily build a VIB. This study refers to APM to build the basic element of figural testing and transform the item combination rule of APM into image processing actions to be into the VIB for final test and validation.

The result of the test shows a positive correlation with that of using APM and demonstrates a desirable correlation coefficient(r = 0.683, n = 301, p = .000). The item exposure rate was extremely low with the rate ranging from 0 to $1.0128e^{-4}$, while the item overlap rate was $2.43488E^{-10}$which could be excluded from calculation. Conclusively, when VIB is used, test security is the highest and an examinee's ability can be correctly calculated.

Above all, the VIB building process proposed in this study are well-acclaimed by both test editors and experts. We can use this technique to build a VIB on all tests. With regards to research tool manipulation, both test operators and experts involved in this study think the tools are easy to use. Using graphic design technique to build objects and rules make it easy to build a VIB. On the test interface, CAT can quickly generate an item. Besides, both test operators and experts have not seen any duplicate items during this study, which means test security was ensured along the way.

Copyright © The Turkish Online Journal of Educational Technology

SUGGESTION

However, this study also met some limitation along the way. In terms of developing research tools, for example, some test editor thought that they are difficult to input other item types. Also, some rules, such as four arithmetic operations, cannot be correctly loaded into the system. Besides, the problem-solving rules and item composition rules of some tests are extremely complicated as they need more time, manpower, and resources to be loaded into the system than designing a test. In an English grammar test, for example, we should take into consideration the item composition rule, but also should make sure the whole context is meaningful. However, VIB can't check if the whole context is meaningful. Another example is Cube Counting testing. Cube Counting testing is extremely complex as it should take into consideration the angles that human eyes cannot see. Therefore, items generated by a VIB may not be solvable.

In terms of difficulty assessment, some test editors found it difficult to assess the difficulty of a test. Because Items of similar types has different levels of difficulty, there is still room for improvement when assessing the difficulty of a test. It is even challenging to assess the discrimination parameter and guessing parameter. Besides, it is more difficult to assess the discrimination parameter and guessing parameter of VIB than a traditional item bank.

In terms of examinees, some examinees found distracters too difficult. As distracters were generated using content-based image retrieval technique, some distracters were so similar that examinees made misjudgments and their scores were affected. Besides, sometimes item variety can be extremely small to mislead the examinees. Some examinees suggest VIB be used in practice systems because there are almost no item exposure problems in VIB. Some examinees have shown improvements of some extent in their grades through extensive practice.

In this study, we have seen great results in tests with simple objects and easy problem-solving technique. This study will take a different approach, such as human intelligence and fuzzy computing technique, to solve the above-mentioned problems concerning research tool development and difficulty assessment. As human intelligence evolves over time, we hope that fast-speed computing can easily solve all kinds of problems in test theories. This study will also plan to use VIB in all types of tests so that we can find flaws in VIB and correct them to ensure a safer and more efficient VIB.

REFERENCES

Arendasy, M., Sommer, M., & Ponocny, I. (2005). Psychometric approaches help resolve competing cognitive models: When less is more than it seems. *Cognition and Instruction*, *23*, 503-521.

Bejar, I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development*, 199-217.

Berretti, S., Del Bimbo A., & Pala, P. (2000). Indexed retrieval by shape appearance. *VISP'00,147*(4), 356-362.

Brown, L., Sherbenou, R., & Johnsen, S. (1982). *Test of Nonverbal Intelligence: A language-free measure of cognitive ability*. Austin, TX: Pro-Ed

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404-431.

Chang, S. W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *40*, 71-103.

Chen, S. F. (2007). Degree in computer-based test of Taiwan retrospect and prospect. *Education Research and Development*, *3*(4), 217-248.

Chen, S. Y. (2004). Computer adaptive testing questions exposure control methods of test. *Technology and ability to assess indicators International Symposium*, *0*, 3.

Chen, S. Y., & Lei, P. W. (2005), Controlling item exposure and test overlap in computerized adaptive testing, *Applied Psychological Measurement*, *29*, 204-217.

Chen, S. Y., & Lei, P. W. (2010). Investigating the relationship between item exposure and test overlap: Item sharing and item pooling. *British Journal of Mathematical and Statistical Psychology*, *63*, 205-226.

Chen, S. Y., Lei, P. W., & Liao, W. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *61*, 471-492.

Chen, S. Z. (2007). *Competency-based distribution of SHC exposure control method*. Unpublished master dissertation, National Taichung University of Education, Taichung, Taiwan.

Cronbach, L. J. (1990). *Essentials of psychological testing. (5th edn. ed.)*. Harper & Row, New York.

Davey, T. & Parshall, C. G. (1995). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, USA.

Drasgow, F., Luecht, R., & Bennett, R. (2006). Technology and testing. In R. Brennan (Ed.) *Educational Measurement 4th Ed*. Wesport,CT: NCME/ACE.

Copyright © The Turkish Online Journal of Educational Technology

Euripides, Petrakis, G.M., & Evangelos M. (1999). Efficient retrieval by shape content. *ICMS'99*, *2*, 616-621.
Finlayson, G.D., Chatterjee, S.S., & Funt, B.V(1996). Color angular indexing. *ECCV'96*, *11*, 16-27.
Freund, Ph. A., Hofer, S., & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*, *32*, 195-210.
Ho, R. G. (2000). *Computerized tests and some related issues*. Paper presented at GCCCE 2000 (The Global Chinese Conference on Computers in Education 2000). Singapore.
Lai, H. , Alves, C., & Gierl M. J.(2009). *Using Automatic Item Generation to Address Item Demands for CAT*. Presented at the CAT Research and Applications Around the World Poster Session.
Mehtre, B. M., Kankanhalli, M. S., & Lee, W. F. (1998). Content-based image retrieval using composite color-shape approach, *Information Processing & Management*, *34*(1), pp. 109-120
Millman, J., & Westman, R. (1989). Computer-assisted writing of achievement test items: toward a future technology. *Journal of Educational Measurement*, *26*(2), 177-190.
Proceedings of the 27th annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.
Singley, M., & Bennett, R. (2002). Item generation and beyond: applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.). *Item generation for test development*, 361-384.
Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing*. (Research Rep. 95-25). Princeton, NJ: Educational Testing Service.
Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*, 57-75.
Sympson, J. B., & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing*.
Tsai, Y. C., & Kuo, B. C. (2007). *The investigation of CAT with item exposure rate controlling*. Unpublished master dissertation, National Taichung University of Education, Taichung, Taiwan.
Tsai, Y. F., & Chen, S. Y. (2008). *Compare online computer adaptive testing test overlap rate control method*. Unpublished master dissertation, National Chung Cheng University, ChiaYi, Taiwan.
Yu, X. J., & Wong, C. S. (1990). *Raven's Progress Matrices guide*. Taipei: Chinese Behavioral Sciences Society.
Yu, M. N. (1993). The introduction of the item response theory (XI): the establishment of Exam. *Study Information*, *10*(4), 9-13.

Copyright © The Turkish Online Journal of Educational Technology