Framework and Implementation of the Virtual Item Bank System

Wen-Wei Liao[1,2] , Rong-Guey Ho[2]
Information Management Department, Chinese Culture University, Taipei, Taiwan[1]
Graduate Institute of Information and Computer Education, National Taiwan Normal University, Taipei, Taiwan[2]
Email: abard@ice.ntnu.edu.tw, hrg@ntnu.edu.tw

***Abstract*- Bagua, I Ching was applied to tell the fortune of people in ancient Chinese. In the modern era, we apply tests to infer the intelligence, future development direction and potential of people. However, it is not easy to design tests, and the security issue has also become a difficulty for test designers. This study has employed Item Response Theory (IRT) and Content-based image retrieval (CBIR) to establish an item bank. There are no actual items in the item bank. What replaces it is a Virtual Item Bank system (VIBs). In the VIBs, there are only the basic objects and processes in the VIBs. The items that are created by adopting the systems are directly created through the objects and processes. The system completely resolves the security issue of the item bank, and a variety of exercise systems created by adopting the system also have considerable help in enhancing students' abilities.**

***Index Terms*—CBIR, IRT, VIBs**

I. INTRODUCTION

Along with computer technology being widely applied in teaching, adopting computers to process tests is already an important trend. ETS (Educational Testing Services) has promoted Computer Based Training (CBT) from 1990. For example, GRE (Graduate Record Examination) has processed tests with CBT from 1992, and from 1993. IRT is also combined and tests are implemented in Computerized Adaptive Testing (CAT). TOEFL (Test of English as a Foreign Language) computer version started to be implemented from 1998, and Taiwan also started to apply CAT from 2000 (TOEFL-CBT). ETS changed TOEFL-CBT to TOEFL-IBT from 2006, and the old computer TOEFL was then put out of action [1].

The greatest difference between CBT and CAT is that CAT will alter along with the previous question's answering status of the test taker immediately, the entire test is specially designed according to the test taker's ability and skill, which is, according to the different abilities of the test takers, different questions will be offered. In short, if the test taker answers the first question correctly, the second question will be harder. On the other hand, if the test taker answers the question incorrectly, then the second question will be easier. During the process, the difficulty level will be adjusted according to the answering status of the test taker to select the questions that are most suitable to the test taker's current ability, and the process will be repeated until the predetermined standard is achieved (or the measurement error is within the tolerance level).

As a result of the reduction in both testing time and testing items, many studies have since focused on the application of CAT [2]. Nevertheless, the problems associated with the development of item bank still remain unresolved, primarily due to manpower, budget and time constraints.

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 [3]. There are at least eight figural test development steps, including designing test specifications, editing items, collecting pre-test data, analyzing items' parameters, revising items, selecting an appropriate scoring method, formal testing, and assessing the overall success of the test.

Item exposure rate is one of the most important factors that influence the security of a figural test. The most common way of reducing this risk is to impose a maximum exposure rate. Several other methods have also been proposed in line with this aim [4] [5]. All of these methods establish a single value of r throughout the test. In this study, we present a new method, known as the Virtual Item Bank (VIB) method, which creates an item bank with unlimited items. We will attempt to describe the implementation of VIB and evaluate its' performance with an empirical experiment. In this way, item exposure rate is always 0. Hence, the problems associated with item exposure can be resolved.

II. LITERATURE REVIEW

The study develops the virtual item bank system by referring to the relevant studies of IRT, CAT, DATA Mining, and the automatic item-generation system in computer-based figural testing. The related literatures are as follows.

A. Item Response Theory

IRT is a series of mathematic models mostly used to analyze the scoring of tests or questionnaire data. The objective of these models is to determine if the latent trait expressible through test. These models are currently used extensively in psychological and educational measurements. IRT was developed in 1960s by Danish statistician, Georg Rasch, and American psychological statistician [6], Frederic M. Lord [7], simultaneously in their respective country. Despite of the different study approaches applied, their results were quite similar.The IRT model is as in below:

doi:10.4304/jsw.7.6.1329-1337

$$P(\theta) = c + (1-c)\int_{-\infty}^{a(\theta-b)} \frac{e^{\frac{-t^2}{2}}}{\sqrt{2\pi}} dt \quad \text{.....(1)}$$

This model is named as 3-parameter Normal-ogive model (3PN) by Lord. To practically simplify the numerical treatment, 3-parameter Logistic model (3PL) is used more often. The model is as in below:

$$P(\theta) = c + \frac{(1-c)}{1+e^{-Da(\theta-b)}}$$ D is the constant 1.7……(2)

The curve based on these two models is Item Characteristic Curve (ICC), which describes the relationship between "the possibility of successfully solving a specific item in the test" and "examinee's ability" (which is denoted as θ in the function). There are three parameters in above two models, a, b and c. Parameter C is named as the guessing parameter. As indicated below, c represents the lower limit of ICC, meaning intuitively that c is the guessing ability, the probability of an examinee making a good guess even though his ability is extremely low, closing to negative infinity.

b is named as item difficulty. B is the value of θ at the point the maximum slope on the ICC. To ICC with a lower limit of 0, b stands for an examinee's ability at the probability of 0.5. The change in B leads to a shift of ICC to either the right or the left without altering its shape. For example, a decrease in the value of b leads to a left shift of ICC, meaning that the test becomes easier.

a is the item discrimination. The value of a/4 is the maximum value of slope. A minor change in the value of ability leads to a hugest change in P at this point.

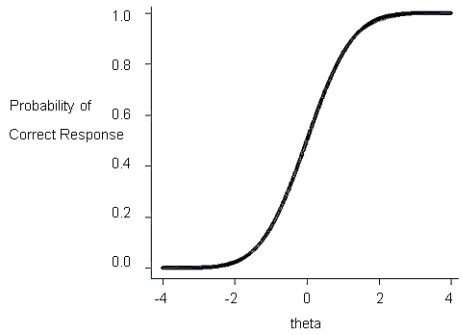

Figure 1. ICC Function

The model proposed by Rasch is as in below:

$$P(\theta) = \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}} \quad \text{.....(3)}$$

It is a multiplicative gamma model for reading speed originally proposed by Rasch (1960) in his monograph Probabilistic Models for Some Intelligence and Attainment Tests [8]. Some IRT researchers regard Rasch's model as a special case of the 3PN model with the parameters c and a at the value of 0. Others consider Rasch's model completely different and that the model really demonstrates the definition of "measurement" because θ and b were defined respectively as "the number of correct responses" and "the correct response rate for a specific item" when the model was proposed. Besides, Rasch's model is more concise.

B. Computerized Adaptive Testing

In this research, CAT theory was applied in the CAT system, turning measurements into tailored tests. CAT is very different from traditional tests because it selects the most appropriate items for examinees based on their abilities or characteristics. If an examinee gets a right answer, a more difficult item will be selected; on the other hand, if examinee gives a wrong answer, an easier item will be asked. Item Response Theory (IRT) just provides serious concept foundation for CAT.

In general, CAT procedures include three important parts: test starting and ending points, ability evaluation and item selection, and the result [9]. After determining the starting and ending points, it begins with the first item; after receiving the answer, it evaluates the ability of examinees and selects the most suitable questions for the next item until it reaches an ending condition. The flow chart of the item selection and ability evaluation is shown in Figure 1, and the followings are the discussions.

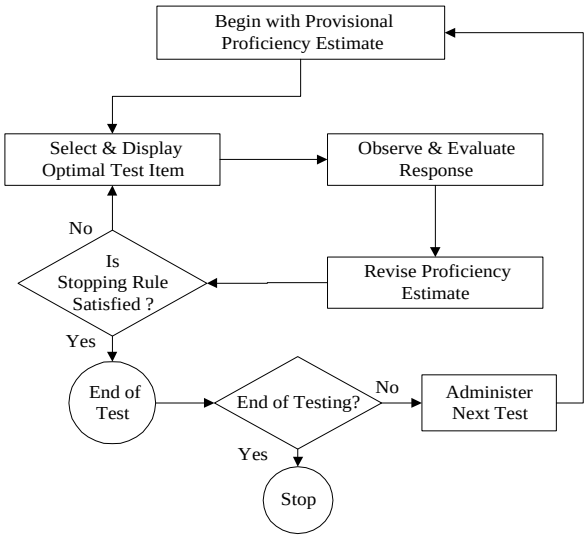

Figure 2. Flowchart of Computerized Adaptive Testing

The selection of starting items in CAT is very important because a suitable beginning item can decrease the length and time of the tests. There are three methods of determining starting and ending points for common multiple choice tests:

(1) Medium difficulty item: in general, examinees are in the medium level; thus it can be assumed that the ability of examinee is of average degree, and the system can be started by selecting median difficulty items.
(2) Random selection: the computer can randomly select items, their difficulties are between -0.5 and +0.5.
(3) Examinees' data: the computer determines the starting and ending points according to examinees' age, intelligence, grades, characteristic, and other data.

In this research, the medium difficulty item was selected for the starting points. Items being generated by

the medium difficulty item generation rules were the first items in the CAT system.

The step of Ability Estimation and Item Selection was a recursive process for the ability estimation and selection. After examinee answered the questions, his or her ability would be re-estimated, and the item selected based on the estimation until the ability evaluation was accurate. The most commonly used ability estimation methods were the MLE and Bayesian Model in IRT; and the item selection strategies were Maximum information strategies and Bayesian strategies [10].

MLE was easier in terms of ability estimation. It could estimate the examinees' ability accurately when the number of items was sufficient; however, if the examinees appeared abnormal (e.g. getting all the answers right or wrong answers), it would not end [11]. The formula is:

$$\theta_{m+1} = \theta_m - \frac{[\frac{d}{d\theta}\ln L(u \mid \theta)]_m}{[\frac{d^2}{d\theta^2}\ln L(u \mid \theta)]_{m+1}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

θ: The ability of the examinee

u: The response pattern of the examinee, u = 1 means the item response is correct, and 0 means incorrect answers

The Bayesian Model assumed that the ability posterior chance ratio was the product of the maximum function and prior ability.

$$posterior \propto likelihood \times prior \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

It could prevent from being unable to end. But the efficiency was lower than that with MLE, and it had a regressive effect, which might lead to deviation [12].

In terms of the selection strategy, Maximum Information Strategies was commonly used. Since information amount and test deviation were negatively correlated, the same item provided different amount of information to examinees with different ability; and different items provided various information amount to examinees with same ability. Thus, the selection should be below the ability of the examinee, and the item that provided the most information should be the next item. The above mentioned was the principle of Maximum information strategy.

In CAT, different examinees have differentiated test length. In general, there are three methods to end the tests.

(1) Set the maximum number of items, namely, preset the test length. After the examinee finishes the maximum number, the test is over.
(2) Set the minimum error standard: when the examinees' ability deviation is lower than the minimum deviation, it means the ability estimation is stable, and leads to the end of the test.
(3) No more suitable items in the item bank: if none of the items can provide more information, it means the additional item does nothing to the ability estimation. Then, the test is over.

In this study, the Bayesian method was used to evaluate the ability, and the Maximum information strategy was used to select the next item.

C. Relevant studies of the automatic item-generation system for computer-based figual testing

Computer-based figural testing has been widely employed across various institutions, such as the Online Testing Center (http://www.onlinetest.org/), the center of Applied Psychology at Beijing Normal University (http://www.bnufr.com), and commercial web sites like IQTest (http://www.iqtest.dk/). These organizations provide useful computer-based figural testing tools and analytical (analysis) tools for researchers. However, only online versions are provided.

Lin (2001) has researched computer adaptive figural testing since 1998 [13]. His researches are based on the analysis of Raven's Advanced Progressive Matrices (APM) test structure, besides being responsible for the development of the New Figure Reasoning Test (NFRT). NFRT contains two main systems: the automatic item-generation system and the online testing system. The online testing system based on IRT theory is just an interface for collecting and evaluating the ability of examinees. The point of this study is an automatic item-generation system which will be discussed in the following paragraphs.

An automatic item-generation system contains an item generation algorithm and an item-generation engine based on APM. The functions, strengths and restrictions of this system are described as follows:

(1) Item generation engine: The engine can automatically generate a specific item with particular content features, and combine different types of geometric figures in a systematic fashion for producing and measuring the item which matches the goal. The purpose of the measurement was to evaluate examinees' reasoning ability on the conclusion (inference on relations) and deduction (inference of relativity) through the figure partition characteristic of the item and the manipulation of the relationships between figures in space. An example item of APM is shown in Figure 3.

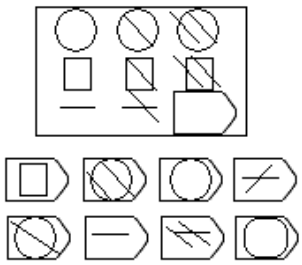

Figure 3. An example item of APM

(2) Item generation algorithm: The algorithm for item-generation was based on the understanding of the analysis of features in APM items. The key points were the parameters in IRT theory and the problem solving processes of APM.

The IRT parameters of APM are discussed as follows:

(1) Difficulty: According to Hambleton and Swaminathan (1985), the value of item difficulty parameter was set to between -2.0 and 2.0. Based on this criterion [4], the average difficulty of APM items was -0.868, and between -2.0 to 2.0.
(2) Discrimination: In terms of ability tests, the value of discrimination paratemter was more than 0 and relatively low in APM, and the item 8 had the lowest discrimination (0.014).
(3) Guessing: According to the estimation, the supposed value in APM items was 0.219. Since there were 8 choices in APM, the predicated value should be 12.5%. The average guessed value was higher than expected.

D. Relevant studies of selection verification

In selection verification, what the test administrators care the most about is the accuracy and difficulty level of the options. Figural testing verification is much more difficult than text testing. While the multimedia science becomes more and more developed, the study employs the technology of content base image retrieval for selection verification. The related technologies are as follows:

(1) function without color characteristics:
The simple eigenvector fl can be used to represent the figure when no color features are considered.
$f_i = (i_1, i_2, i_3, \ldots i_n)$…..……..……..……..(6)
f is the eigenvector of figure i, and n is the content feature number. The similarity level of the two figures calculates the Euclidean distance of the eigenvector (as shown in function 7). The closer the value is to 0, the higher the similarity level of the two figures is. The greater the value is, the lower the similarity level is.
$d(Q,I)=\sqrt{\sum_{j=1}^{n}(f_j^Q - f_j^I)^2}$ ……(7)
(2) function that consider the color features:
If colors need to be considered, then other methods must be employed. Mehtre, Kankanhalli and Lee (1998) proposed to comprehensively consider the two eigenvalues of figure color and shape, calculate the similarity level of the figure, and apply logo comparison as the study subject. The steps of the proposed method are as follows:
(I) Look for the color clusters in the figure. The calculation of the color distance is shown as function (8). While processing clustering on the 400×400 figure color used in the experiment, the color distance minimum threshold between each clusters is set to be 50.
Color distance=$\sqrt{(\Delta R)^2+(\Delta G)^2+(\Delta B)^2}$ ……..(8)
(II) Look for the shape clusters in the figure, first divide the classified color clusters into several layers according to step (I). The color cluster number is the layer number. Mark the shape clusters of each layer figure, and sort the shape cluster, in descending order, on the figure of each color layer according to the pixel amount of each shape clusters. When the point in the shape cluster is less than 50 pixels, the shape cluster can be ignored. In addition, to avoid false determining a thin line as a cluster, the minimum density value of the shape cluster (see function 9) is set as the threshold of various shapes. If it is smaller than the value, then the shape cluster should were ignored.

density $\rho = \frac{population\ of\ Cluster}{(l_{\max})^2}$ ……..(9)

$l_{\max}$ =max (||x_2- x_1||, ||y_2- y_1||)

(x_2, y_1) and (x_2, y_2) are the corner points of the shape cluster.

(III) Similarity level calculation:
Respectively calculates the similarity level of color and shape according to the calculation function of color and shape distance (see function 10 and 11), and then calculates the similarity level of the two integrated features according to function 11.

$coldis(C_i^Q, C_j^I)=\sqrt{(R_i^Q - R_j^I)^2+(G_i^Q - G_j^I)^2+(B_i^Q - B_j^I)^2}$ ……(10)

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

$shpdis(C_i^Q, C_j^I)=\sqrt{\sum_{i=1}^{7}(m_i^Q - m_i^I)^2}$ ……... (11)

I is the moment invariants

$D(Q,I)=\omega_1\psi_1+\omega_2\psi_2+\omega_3\psi_3+\omega_4\psi_4$ ……(12)

$$\psi_1 = \sum_{i=1}^{\max(m,n)} cdist(C_{c,i}^Q, C_{c,Pc(i)}^I)$$

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

$$\psi_3 = \sum_{i=1}^{\max(p,q)} shpdist(C_{cs,i}^Q, C_{cs,Ps(i)}^I)$$

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

,

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

Pc is the closest color cluster assignment function, it can maps every color cluster i of image Q to the closest color cluster Pc (i) of image I.

III. METHODS

The objective of this study is to propose a new concept – VIB, and to show how this concept is used in the CAT. The following is a discussion on the problems and the demands of item bank generation we encountered in addition to the development of the research tools.

A. Problems and demands of the item bank generation

The item bank consists of calibrated, analyzed, categorized, and evaluated items. Millman and Arter (1984) believed that items would be more computerized in the future. The IRT-based item bank was estimated to have the following advantages:

(1) It allowed test editors to edit items for all purposes without any restraints.
(2) It allowed test editors to edit tests with the proper amount of items in the range of the item bank [15].

Thus, an item bank has the potential to improve the test quality. However, we often face the following problems while building an item bank:

(1) Number of items:
In general, it is better to have more items. But it should be taken into consideration whether the item's quality has reached the test editors' requirements as well as achieving the purpose of the test. Researchers suggested that every concept must include 10 items, and every course unit had to contain 50 items. Reckase (1981) recommend 100 to 200 difficulty parameters distributed evenly, and items with the discrimination parameter. If this standard could be reached, it could be used for computerized adaptive testing [16].
(2) The categories of the item bank:
The most common categorization is one using the theme or instructional goal, and the other is using key words to search. In general, using key words is more flexible and could be used for certain purposes, content, age, and thinking style.
(3) Scaling parameters of items:
Scaling parameters are designed to calibrate item parameters like difficulty, and convert them to the same scale. In the test of a large sample, scaling parameters are necessary; however, they could be omitted for individual tests.
(4) The problem of public access:
It seemed that the teaching might be limited to the content of the item bank if the teachers use the item bank as assessment tool freely. But if the item bank was large enough, this problem could be ignored because teachers were unable to limit the teaching to the item bank content. On the other hand, if the item bank was not large enough, opening the item bank might lead to narrowing the focus of teaching. Thus, it must be considered whether the item bank should be open or not. But, opening a few item samples could help both teachers and students to understand the testing method, something both necessary and correct.
(5) Security problems of the item bank:
Item banks could make test editing and scoring easier; however, it requires repetitive use of the item bank and can interfere with the item security (such as through appearance of old items). This has to be taken into consideration if item bank is small; on the contrary, this concern could be ignored if the item bank is large enough. In addition, updating the items constantly to ensure content validity and statistical quality is another way to ensure the item bank security.

Based on these considerations, we found that a test with a sufficient number of items could be helpful in quality, safety, and teaching. Thus, this study expected to design a new concept for an item bank containing an abundant numbers of items with the fair quality to solve the problems mentioned above.

B. Development of the research tool

This research has developed two research tools: Virtual Item Bank System and CAT system. The system structure and functions of these two tools are described as follow.

(1) Virtual Item Bank System (VIBS): In VIBS, the item database no longer stores large amounts of items; instead, it saves two elements to replace the traditional items:
(I) Basic figure object: This system no longer requires saving a large amount of figural items. Instead, items were built upon three basic figure types: line, circle and multilateral. Not only does this lower the memory space requirements, but it also reduces the probability of item exposure.
(II) Processes: The examinees' solving processes and abilities were defined by specialists and converted to mathematical formulas which could be manipulated by computers and stored in the hypothetical item database. Using this data along with the basic figure objects, the computer can produce mass items and lower the work load for test preparation.

The VIB which replaces the traditional item bank is illustrated by the flow chart below.

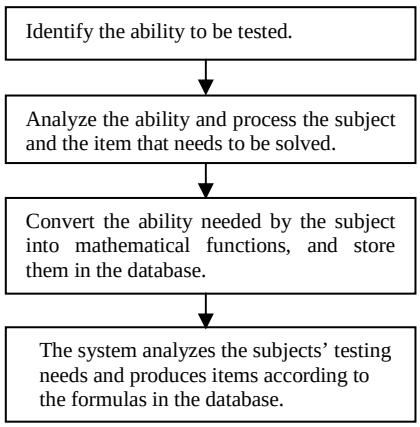

Figure 4. Flow chart of VIB

The VIBs contains three subsystems: item rule definition subsystem, item generation subsystem, and answer retrieval subsystem. Each subsystem has different tasks and functions and is described below.

(1) Item rule definition subsystem:
This subsystem provides test editors with a number of figural objects, processes the needed information to solve the problem. Through the system interface, users can determine the figure's position on the system's interface and choose the method to process images. The subsystem then estimates the item difficulty and asks the test editors to adjust the difficulty level. Finally, the item initiation subsystem would save this information into the data.

Referring to the terms of parameter estimation (with respect to the IRT parameters), test editors need to consider examinees' experience, required ability, age and other factors since these are a number of the strands that could affect item difficulty.

The study expected to help test editors to automatically define the difficulty level of items, and lessen their burden. This study has simplified most factors while deriving parameters and analyzing the amount of objects needed (items needed) and image processes. Other factors will be analyzed in the later sections.

In terms of the parameter estimation, there are three methods based on various parameter conditions:

(I) If the item parameter is already known and only ability parameter is needed to be estimated: the MLE and Bayesian procedure were commonly used [17].

(II) If the ability parameter is known, but we need to estimate item parameter: we used MLE and Bayesian Procedure [17].

(III) If the item and ability parameter were both unknown: we used Joint Maximum Likelihood Estimation (JMLE), Marginal Maximum Likelihood Estimation (MMLE), Bayesian Model or Maximum a Posteriori Estimation (MAP), Bayesian Mean or Expected a Posteriori Estimation (EAP) to estimate item and ability parameter [18].

In this study, the ability parameter was used to estimate since ability parameters were unknown.

(2) Item generation subsystem:

The main function of this system was to generate all kinds of data in the item generation subsystems in the hope of producing an infinite number of items. It contain the main functions of the item generation system are:

(I) Defining the needed abilities and strategies in order to solve the item.

(II) Determining the object shape of each item

(III) Identifying the difficulty parameter

(IV) Parameter conversion : The system converts the data mentioned above into mathematical formulas, and saves them in the VIB.

(V) Automatic generation: The item generation subsystem can automatically generate items according to the defined strategy, difficulty level, and selection.

(3) Answer 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 function 13) of the characteristic value, and we selected the lowest three as the alternative option.

$$d(Q,I)=\sqrt{\sum(f^{Q}-f^{I})^{2}} \quad \text{.....(13)}$$

The VIBS is composed of those subsystems which control the item shape, item difficulty, answer, and all parameters.

C. System interface

The system interface and function are as follow:

(1) Decide the location of figural objects

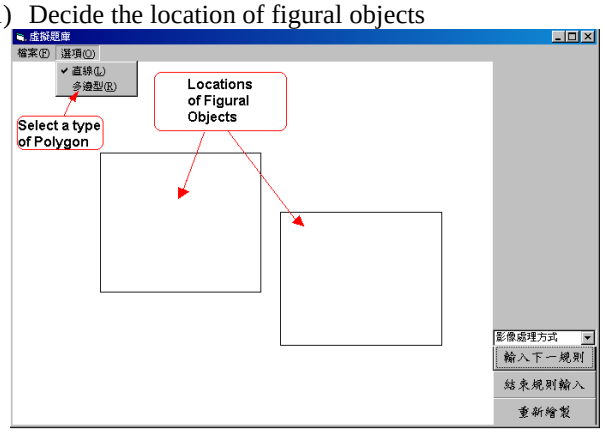

Figure 5. The interface of decide the location.

(2) Decide the processes of objects

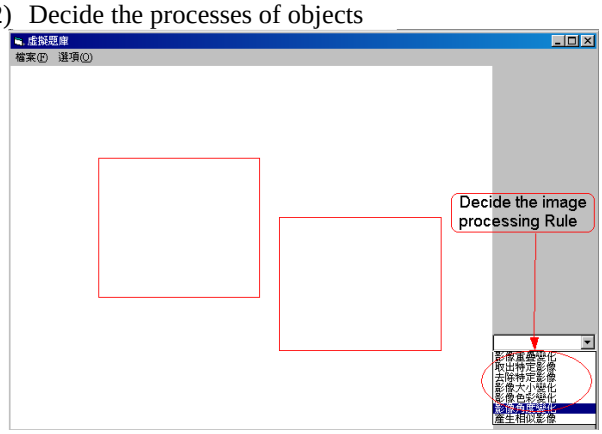

Figure 6. The interface of decide rules (processes)

(3) Step 3: Choose the next figural objects and save them into the VIB.

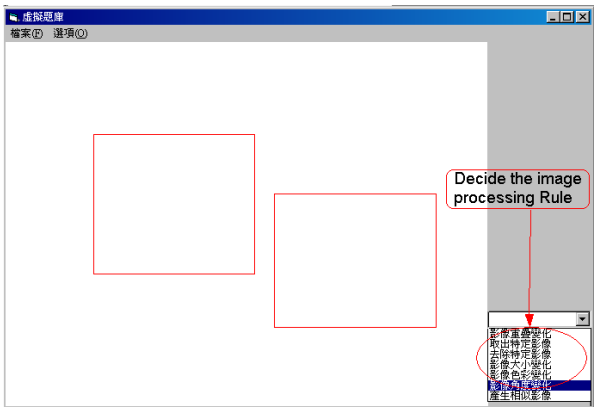

Figure 7. The interface of save function

According to the abovementioned Processes and Element definitions, the system will save it as an XML

file, and when reading VIBS through CAT, the following tests can be generated.

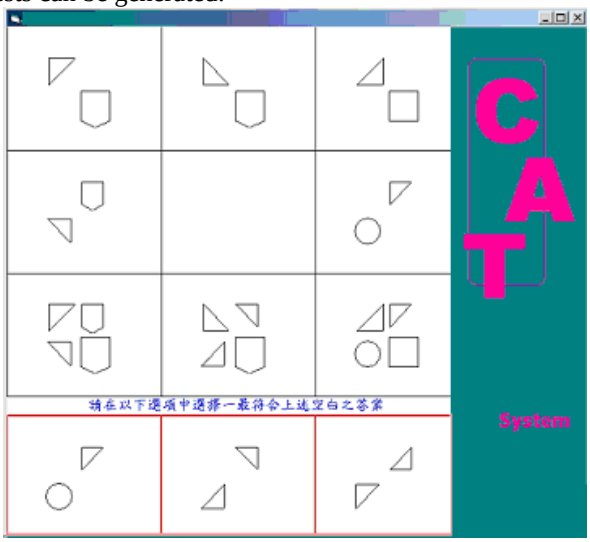

Figure 8. The demo of Computer figural test with VIBs

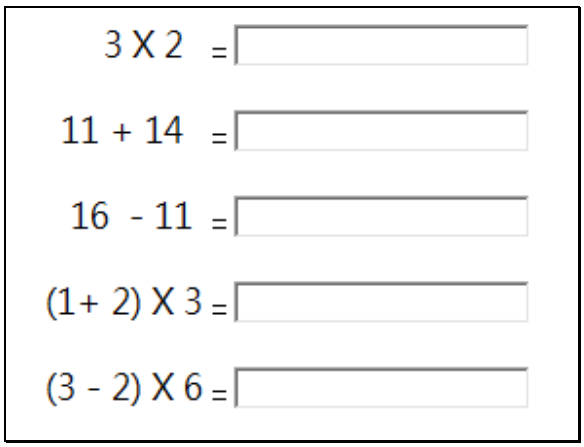

Figure 9. The demo of Four Arithmetic Operations Test with VIBs

Figure 10. The demo of Cube counting Test with VIBs

The figure above represents the issues of the problem and demands of item bank generation, in addition to the development of research tools. The research tools helped test editors to solve the problem of the item exposure rate. A simulation of the item overlap rate will be discussed and proved in the following section.

IV. RESULTS

A. System implementation

The study system employs the Internet 3-tier (Browser-WEB-Application) client/server architecture according to the study purpose and system analysis and design. The following respectively describe the tools and technologies adopted in the system implementation, database design, and system program verification:

(1) Development tool technology: C#, XML

(2) Database design architecture:

The system adopted the XML file format as the backend database. Automatic generated test questions and the functions of adding, searching, amending, and deleting related information can be achieved through the operation of the XML format. The XML format can be coupled with various server operating systems and Web servers, and is suitable as system backend storage tool.

(3) System Algorithm

The following are algorithms for the Item Initial System and Item Generation System.

(I) Item Rule Definition subsystem：

This system is mainly based on the binary operation in image processing theory. When the test editor defines the figure object's location and image processing operations, the System will generate the result and storage it into Virtual Item Bank. The pseudo code for system algorithm is as follow:

INPUT:
1. define objects as {line,circle,polygon}
2. ni [xi.. xj,yi.. yj] n belons to {line,circle,polygon} i belons to {1..10}.xi,xj,yi,yj belongs to {1..7500 (pixel)}}
3. Oi belonging to image process operations {Or, And, Xor, Sub, Color, Size}
4. Pi belong to difficult parameter {-3..+3}

OUTPUT:
1. ri [xi.. xj,yi.. yj] r belonging to {line,circle,polygon} i belonging to {1..10}. xi,xj,yi,yj belonging to {1..2500 (pixel)}}
2. F [i] F is item generation function which is storage in Virtual Item Bank.
3. STEPS
Get the location of normal figures nj (xi,yi)
Get the image process operations oi
Get the location of figures which want join image process nk xi,yi)
Get the difficult parameter pi
For every j belonging to n :
For every i belonging to o :
rj=nj oi nk
function:F (j)= Rj and pi

(II) Item Generation System

INPUT :The locations of Figure Objects, Difficult parameters.

OUTPUT: Figural Items.
STEPS : (initializations).
Pi=Difficult parameters of Itemi
ri=location of final figures
Random select a type of polygon. Si
If Si ever be selected then
Begin
Record the type of Si
  Random select another Si
End
Random select a item shape from Rule
Random select a direction belonging (Top to Down, Down to Top, Right to Left, Left to Right)
Generate the item
Generate the answer
Doing answer Image Data Retrieval
Get the perfect answer

B. Test Security

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 = 12, subjects = 30000 , number of objects = 1, process of the item generation = 12 , the simulation results are as follows.

$$R_t = \frac{T_O / C_2^N}{(\sum_{i=1}^{N} L_i) / N} = \frac{2T_O}{(N-1)\sum_{i=1}^{N} L_i} \quad \text{.....(14)}$$

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 1. Results of the item overlap rate simulation

Item overlap rate (R)	1.714321^{-10}
Mean of test length	9.3012
Mean of Theta-Estimated	-0.134
Mean of SE	0.3017

Table 2. Use frequency of each item-generation rules

Rule	frequency	Rule	frequency
1	19321	7	18765
2	23012	8	17862
3	17632	9	19122
4	18453	10	17280
5	19865	11	22009
6	20121	12	21776

Table 3 Item overlap frequency (times) of each rules

Rule	frequency	Rule	frequency
1	0	7	0
2	0	8	0
3	0	9	1
4	1	10	0
5	1	11	0
6	0	12	0

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

V. Discussion and Conclusion

From the results of item overlap simulation, it is obvious that the VIBs can resolve the problem of item exposure efficiently. Every examinee got different items on the same test. This allows the VIB to be used not only in measurement but also in practice. The results of the experiment showed its evident effects in practice.

In the VIB, the item was generated dynamically. It was however difficult to apply it in the CAT system. In order to solve this problem, two CBT testing systems were designed to collect the item difficulty parameters of the item generation rules.

The study has also encountered some problems. For example, in study tool development, some test designers think that it is difficult to operate, and the method of the questions cannot be correctly entered into the system, such as pentomino. Some problem solving and test combination methods are extremely complicated and the human and material resources cost to input them into the system is even more than designing the test, such as English grammar tests. However, in difficulty estimation, some test designers also found that difficulty estimation is difficult. Some question types are similar; but the difficulty level is completely different, and therefore difficulty consideration is still insufficient.

In addition, some test subjects proposed the issue of the distracters being too difficult. Because some distracters are generated through Image Data Retrieval, some distracters are too similar and result in the test subject making an incorrect decision, further impacting their score. In addition, the change of questions is sometimes very small, which also easily results in the test subjects giving incorrect answers. Finally, some test subjects proposed the recommendation that because VIB will almost not generate question exposure, VIB can be employed in the practice system, and after a large amount of practice, some test subjects' score will make significant progress.

REFERENCES

[1] Yang, H. L., & Ying, M. H. (2005). Could On-line Testing have the Same Effects on Scoring as Paper-and-Pencil Testing? *Journal of Taiwan Normal University: Mathematics & Science Education, 50(2)* , pp. 85-107

[2] Ho, R. G., & Hsu, T. C. (1989). *A comparison of three adaptive testing strategies using MicroCAT*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

[3] Cronbach, L. J. (1990). *Essentials Of Psychological Testing*. New York, Harper Collins Publishers

[4] Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

[5] van der Linden, W. J., Ariel, A., & Veldkamp, B. P. (2006). Assembling a CAT item pool as a set of linear test forms. *Journal of Educational and Behavioral Statistics*, 31, pp. 81–100.

[6] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research.

[7] Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbawn Associates.

[8] Jansen, M. G. H. (2003). Estimating the parameters of a structural model for the latent traits in Rasch's model for speed tests. *Applied Psychological Measurement, 27*(2), 138-151.

[9] Wainer, H. et al. (Eds.). (1990). *Computerized adaptive testing: A primer. Hillsdale*, NJ: Lawrence Erlbaum Associates.

[10] Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.

[11] Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

[12] Ho, R.-G. (1989). Development and implementation of the CAI software database and interactive evaluation system. *Paper Presented at the 1st 1989 International CAI Conference*. Taipei, Taiwan, ROC. (Invited Speech)

[13] Liu, Z. J., Liang, R. K. & Lin, S. H. (2001), 「Automatic Item-Generation and Online Testing System for New Figure Reasoning Test, *5th Global Chinese Conference on Computers in Education / International Conference on Computer-Assisted Instruction 2001*, pp. 326-333

[14] 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

[15] Mehtre, B. M., Kankanhalli, M. S. & Lee, W. F. (1998). Content-based image retrieval using composite colour-shape approach, *Information Processing & Management, 34* (1), pp. 109-120

[16] Reckase, M. D. (1981). Tailored testing, measurement problems and latent trait theory. Paper presented at the annual meeting of the National Council for Measurement in Education, Los Angeles.

[17] Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. NY: Marcel Dekker.

[18] Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*, Boston: Kluwer-Nijhoff.