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.
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 (Simpson & Hetter procedure) proposed by Simpson & 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 (Simpson & 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 a past examinees (α ≥ 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 & Kyellon, 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 test 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...
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:

\[ f_i = (i_1, i_2, i_3, \ldots, i_n) \]

(1)

\[ d(Q, I) = \sqrt{\sum_{j=1}^{n} (f_j^Q - f_j^I)^2} \]

(2)

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:
Mehtrre, 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.

\[ \text{Color distance} = \sqrt{(\Delta R)^2 + (\Delta G)^2 + (\Delta B)^2} \]

(3)

**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 shape cluster as the shape threshold is set.

\[ \text{density} = \frac{\text{population of Cluster}}{l_{\text{max}}^2} \]

(4)

\[ l_{\text{max}} = \max(||x_2 - x_1||, ||y_2 - y_1||) \quad (x_2, y_1) \text{ and } (x_2, y_2) \text{ are corner points of shape cluster.} \]

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).

\[ \text{similarity} = \frac{1}{n} \sum_{i=1}^{n} \left( \psi_1 + \psi_2 + \psi_3 + \psi_4 \right) \]

(5)

\[ \psi_1 = \sum_{i=1}^{max(m,n)} \text{cdist}(C_{c,i}^0, C_{c,Pc(i)}^1), \quad \psi_2 = \sum_{i=1}^{max(m,n)} \sqrt{\lambda_{c,i}^0 - \lambda_{c,Pc(i)}^1} \]

(6)

\[ \psi_3 = \sum_{i=1}^{max(p,q)} \text{shpdist}(C_{cs,i}^0, C_{cs,Pc(i)}^1), \quad \psi_4 = \sum_{i=1}^{max(p,q)} \sqrt{\lambda_{c,i}^0 - \lambda_{c,Pc(i)}^1} \]
\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:

![Diagram of the development procedure](image)

**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, Sherbenou & 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.
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.

<table>
<thead>
<tr>
<th>Addition Rule</th>
<th>Diagonal Rule</th>
<th>Oblique Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Addition Rule" /></td>
<td><img src="image2" alt="Diagonal Rule" /></td>
<td><img src="image3" alt="Oblique Rule" /></td>
</tr>
<tr>
<td>Allocation Change Rule</td>
<td>Quantitative Pair Wise Progression Rule</td>
<td>Size Change Rule</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Difficulty Level</td>
<td>207</td>
<td>.3100</td>
<td>.6811</td>
<td>.5101</td>
<td>.0687</td>
</tr>
<tr>
<td>Test Discrimination Index</td>
<td>207</td>
<td>.4121</td>
<td>.9012</td>
<td>.6801</td>
<td>.0872</td>
</tr>
</tbody>
</table>

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:
This system helps test editors to define “objects” and “rules”. 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.

![Select the locations of figural objects](image1)

Figure 3, Defining object position

![Decide the image processing rules](image2)

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

\[ d(Q, I) = \sqrt{\sum (f^Q - f^I)^2} \] ............................................................(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
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, \( \theta \) represents the examinees’ ability; \( p(\theta) \) represents the chances of examinees with \( \theta \) 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.

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.
Table 2, Difficulty parameter estimations of item generation rules

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

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

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

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

<table>
<thead>
<tr>
<th>Rule A</th>
<th>Standard Deviation</th>
<th>Rule B</th>
<th>Standard Deviation</th>
<th>Rule C</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.051218</td>
<td></td>
<td>0.04158</td>
<td></td>
<td>0.082226</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Score</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computerized Figural Testing</td>
<td>38.68</td>
<td>5.450</td>
<td>310</td>
</tr>
<tr>
<td>APM’ Score</td>
<td>29.40</td>
<td>3.620</td>
<td>310</td>
</tr>
</tbody>
</table>

Table 9, Correlated Coefficient of “Computerized Figural Testing” Score and APM’s Score

<table>
<thead>
<tr>
<th>Score</th>
<th>Pearson Correlation</th>
<th>Sig. (2-tailed)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computerized Figural Test</td>
<td>1</td>
<td>.683**</td>
<td>310</td>
</tr>
<tr>
<td>APM Score</td>
<td>.683**</td>
<td>.000</td>
<td>310</td>
</tr>
</tbody>
</table>

**. 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.
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


