

Design a Computer Adaptive Cubic Enumeration Test based on Virtual Item Bank Theory

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—Spatial ability is not only the daily necessity of our life but also an important skill for mechanists, architects, chemists, surgeons, surveyors and cartographers. At present, whether within our country or abroad, the mathematical programs of elementary and junior high schools both list spatial ability as one of the geometry course materials. Spatial ability correlates with the concept of volume, and cubes are the basic units of volume, therefore cubes are frequently used to introduce and explain the concept of volume. Also, Cubic Enumeration Tests are often used to estimate or further one's mental ability and operational skills of 3-D geometric objects.

Cube Enumeration Tests are usually kept confidential and exempt from public access. Ordinary people recognize it only from the limited resource from the practice exams the bookstores offer. The main reason to keep the exam confidential is to maintain the objectivity of the assessment. In this research, we will combine techniques of Virtual Item Bank Theory and Item Response Theory (IRT) to establish a whole set of adaptive Cube Enumeration Tests. Replaced by a Virtual Item Bank system, this exam collection will no longer consists of any substantial forms of the items. Within this system, all data bases will only consist of basic components: cubes. Consecutively, this Virtual Item Bank will no longer encounter the risk of the item exposure due to the fact that the exam questions will be generated by basic components and IRT.

With advanced technology in Computer Science, this research combines IRT, Computerized adaptive testing theory (CAT), Virtual Item Bank theory and related knowledge about Cube Enumeration Tests— designing a whole set of evaluating tools of convenience, in hopes of making some contributions to e-learning.

Keywords—Spatial ability, Cubic Enumeration, IRT, CAT

I. INTRODUCTION

Zhouyi is used to predict one's fate in ancient China and tests are used to evaluate one's IQ, future direction and potential for development in modern days. The present study builds an adaptive cubic enumeration test to evaluate an examinee's spatial ability. Moreover, as there are different combinations of the eight trigrams in Zhouyi, the present study also combines different basic graphic elements for different tests, hoping to estimate an examinee's spatial ability accurately with this item generation system.

II. LITERATURE REVIEW

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 [1], Frederic M. Lord [2], simultaneously in their respective country. Despite of the different study approaches applied, their results were quite similar.

The IRT model is as in below:

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

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:

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

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.

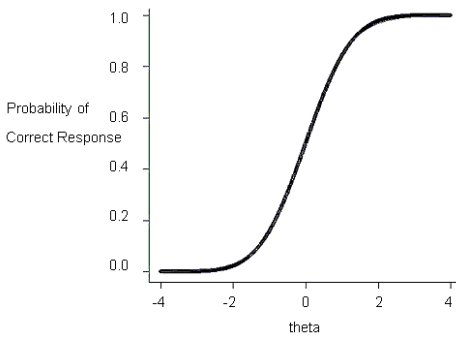


Figure 1. ICC Function

The model proposed by Rasch is as in below:

$$P(\theta) = \frac{e^{(\theta-b)}}{1 + e^{(\theta-b)}} \dots\dots\dots(iii)$$

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 [3]. 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. CAT

A test evaluates an examinee's ability most accurately if the test is difficulty-appropriate for the examinee. Therefore, any test result can't be the most accurate ability measurement for every examinee because the difficulty of the test can't be appropriate for every examinee. An ideal test would be a test adaptive to each examinee's ability and their individual situation.

However, adaptive tests take a huge amount of time and effort, which is no longer a problem after the invention of computer. CAT assigns a question to the examinee according to his response to the last question. CAT assigns the question most informative about the evaluation accuracy of the examinee's ability. As a result, the length of test is shortened. Students with high ability no longer take those easy questions while those with low ability are also free from the difficult questions because they offer extremely limited information for the evaluation of students' ability.

The basic method of adaptive tests includes six steps described briefly in below [4]:

- Select the appropriate IRT model. 3PL model is selected most of the time.
- Select the appropriate item bank.
- Decide which item should be the first item: the common methods to decide which item should be the first item are: random pick from the difficulty-appropriate items, random pick from the entire item bank, survey of

students' background before deciding which type of item should be selected.

- Other frequently used item selection strategies: (i) Selection of the item most informative about evaluation of examinee's ability. To minimize item exposure, Green (1984) suggested a random selection among the most informative items. (ii) Bayesian item selection. The distribution of examinee's ability is set as prior distribution. The number of correct responses, incorrect responses and unanswered items are the posterior variables. The item minimizing the variance of the posterior distribution of examinee's ability is to be selected. (iii) Selection of the item closest to the estimated level at examinee's current stage.
- Scoring algorithm: in the adaptive test, examinee's ability is reevaluated after every new response. Maximum likelihood estimation and Bayesian estimation are the two algorithms commonly used in adaptive tests.
- Termination criteria: the termination criteria for adaptive tests highly relates to item selection and scoring algorithm described previously. If item selection is based on how much information provided by each item, the test is terminated when the accumulated information meets preset criteria. If item selection is based on Bayesian estimation, the test is terminated when the variance of ability estimation is smaller than preset criteria.

C. Automated Item Generation for Matrix Inductive Reasoning Test

Liu (2001) designed the "Automated Item Generation and Testing System for Matrix Inductive Reasoning Test" based on the theory of cognitive measurement approach [5]. The automated item generation in this system consists of an "item-generation engine" and an "item-generation algorithm". The function of the "item-generation engine" is the automated generation of items with certain specific characteristics, a systematically automated combination of all types of item characteristics to generate the items that meet the goal of the test. Its typical item is as in below:

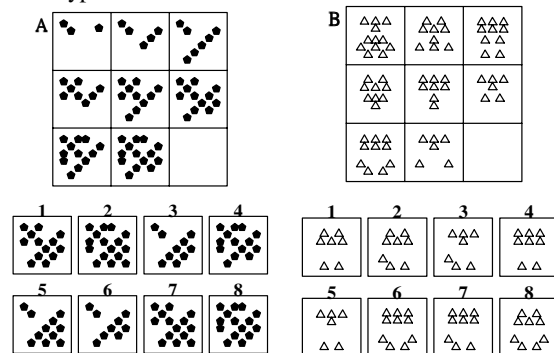


Figure 2. The Typical Items of "Matrix Inductive Reasoning Test"

The stem in the system consists of 3x3 blocks. There is a figure in each block except for the last block in the lower right

corner. Examinees must compare the changes in attribute of each geometric figure (such as number), find out the solution rules and infer the geometric attribute for the missing figure before selecting a proper figure from the options given. The patterns are in the four categories as in below:

- 1) **Constant in row rule:** the same property (such as number) occurs throughout a row with consistent change throughout a column.
- 2) **Quantitative pairwise progression rule:** A qualitative increment or decrement occurs in an attribute of the geometric figure (such as number) between adjacent entries from the left to the right in each row or from the top to the bottom in each column.
- 3) **Addition or subtraction rule:** A geometric figure in the first row/column is added to or subtracted from another figure in the second row/column to produce the third figure in the third row/column.
- 4) **Distribution of three values rule:** the three figures in each row/column consist of different number of geometric patterns while the order of the three diverse figures is different between any two rows/columns.

This research arranged the geometric figures according to these solution rules to measure examinees' ability of matrix inductive reasoning. The items were combination of geometric figures, whereas the category of items with single change in a general matrix inductive reasoning test was neglected. Moreover, the practical use of this automated item generation system is limited due to its complicated item design and the need to apply developer's solution rules.

D. VIBS

The term "Virtual Item Bank" was first seen in the research, "Design a Virtual Item Bank Based on Image Processing Technique", by Ho (2002) [6]. Different from traditional item banks, virtual item banks are not physical lists of items but innumerable items generated via item generation rules, which avoid the issue of item exposure rate in the traditional item banks. In that research, the researcher applied principles of image processing and technologies in graphical test and summarized 12 rules, Addition Rule, Oblique Rule, Quantitative Pair Wise, Progression Rule, Move Rule, Overlap Rule, Angle Change Rule, Diagonal Rule, Allocation Change Rule, Size Change Rule, Straight and Oblique Rule and Bias Rule, Image operation with no changes. The 12 rules combining with the binary algorithm in the four image processing technologies, AND, OR, XOR and SUB, resulted in 48 item generation rules for graphical tests. Infinite and various graphic test items are generated from basic geographic patterns combined according to these rules. Examinees taking a Computer-Based Testing; (CBT) under a virtual item bank system are unaffected by item exposure rate that is a common problem for traditional item banks.

E. Cube Enumeration

Cube, Regular Hexahedron cube, can be used to introduce the concept of volume as the basic unit for volume. Written tasks about enumerating cube arrays are used to measure or

build the mental image and manipulation capability of 3D geometric objects. Students often consider cube enumeration tasks difficult [7], which are quite important in the development of valuable mathematics techniques. A research done by National Assessment of Educational Progress (NAEP) reported that the hit rate for cube enumeration in cuboids was lower than 50% for students at the age of 9, 13 and 17, indicating that cube enumeration tasks, even in regular cuboids, are difficult even for older students [8].

Some researchers applied cuboids consisting of cubes to explore how individual constructed 3D mental objects with 2D figures [7] [8] [9]. These researchers believe that the wrong response in cube enumeration is due to the failed transformation of mental images from 2D figures to 3D mental objects and the lack of perspective to count the hidden cubes.

Cube enumeration tasks for IQ test [10] and mathematics programs are more than regular cuboids and many of the tasks are irregular cuboids. The educational tasks of Battista (1999) and Olkun and Knaupp (2000) are essentially helpful for most students to develop their spatial orientation (spatial structuring), the ability to distinguish between cubes and 2D squares, to build mental 3D objects from 2D figures and to identify the relative location of each cubes.

III. METHODS

A. Research Hypothesis

The present study conducted a preliminary research with a simple written test to six students with diverse levels of mathematics achievements on third grade in the elementary school in April, 2010, followed by interviews with the students about how they arrived at their answer.

The result indicates that students mostly counted the cubes they saw before calculating the number of hidden cubes mentally. Therefore, the present study thinks that what influences the cube enumeration of six-grade students is whether the mental image of hidden cubes can be built. Spatial orientation is influenced by the ability to build correct mental image. The more hidden cubes, the worse performance was observed. The preliminary test indicated that special orientation is one of the major factors influencing the performance of cube enumeration. Therefore, we proposed the hypothesis that the more the hidden cubes are, the more difficult it is for students to build a mental image.

B. System Design

The system developed in the present research is based on IRT three parameter log model. The virtual item bank, IRT testing system and user interface in our system are described in the following.

C. Virtual Item Bank

A virtual item bank is built in two stages as a traditional item bank. The first stage is test construction consisting of three steps, item construction, pretest, evaluation and formal test construction. The second stage is IRT parameter estimation, analyses of examinees responses for item

parameters necessary for the construction of a virtual item bank. The system is show a bellow:

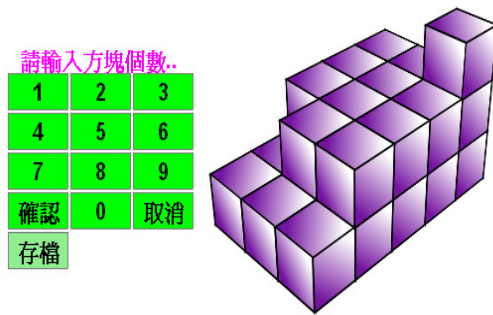


Figure 3. Computer Adaptive Cubic Enumeration Test based on Virtual Item Bank Theory

1) First Stage: Test Construction

- Item Construction:

22 items were constructed as the pretest item bank according to our research hypothesis. Number of different types of items is summarized in the table below.

TABLE 1. ITEMS CONSTRUCTION

Hidden Cubes	Total Cubes	Hidden Cubes	Total Cubes
3	16	5	17
3	16	5	17
3	15	5	17
3	16	6	17
3	16	6	19
3	15	6	17
4	16	6	19
4	17	7	20
4	16	7	21
4	17	7	20
5	17	7	21

- Pretest and Evaluation:

The items were used in a pretest of 231 third-grade students in the elementary school for item selection for the formal test. Items were selected by their respective item difficulty index (P), discrimination index (D) and point-biserial correlation (rpb) that is the product-moment correlation of the total score of 231 examinees. The item selection criteria for the present research are pretest difficulty index is between 0.2 to 0.9, discrimination index is above 0 and point-biserial correlation is significant at the significance level of $\alpha=0.05$.

- Formal Test Construction:

Inappropriate items judged by their item difficulty index, discrimination index and point-biserial correlation were excluded after pretest analyses and an item bank for

the formal test of 20 items were constructed. Number of different types of items is summarized in the table below.

TABLE 2. TABLE 2 ITEMS CONSTRUCTION

Hidden Cubes	Total Cubes	Hidden Cubes	Total Cubes
3	16	5	17
3	16	5	17
3	15	6	17
3	16	6	19
3	16	6	17
4	16	6	19
4	17	7	20
4	16	7	21
4	17	7	20
5	17	7	21

2) Stage Two: IRT Parameter Estimation

The three parameters, item difficulty index, discrimination index and point-biserial correlation, for each formal test item were estimated with the responses of 231 examinees according to IRT three parameter model. Chi-square Goodness of Fit Test was also conducted to test if IRT model fits after parameter estimation. The test result indicated that test data fitted for the IRT three parameters log model applied.

Kuder-Richardson Reliability is applied for the present research based on the test result of 231 students in the pretest sample. The overall α value is 0.9361, indicating certain consistency.

IV.RESULT AND DISCUSSION

Examinee ability was estimated with a computer test taken by 50 examinees randomly selected from the pretest sample with the virtual item bank system developed by this research. To test the capability of the system to estimate examinee spatial ability, a dependent sample t-test was conducted to compare the ability estimation and the true spatial ability measured by the formal test of 20 items.

The result is insignificant with t-value at 1.254 ($p = 0.21 > 0.05$). Therefore, we accept that there was no significant difference between the system-estimated ability and the true value of ability. As the result, the virtual item bank system developed by the present research is verified for its precise estimation of examinees' true ability.

REFERENCES

- [1] Rasch, G., *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research, 1960.
- [2] Lord, F. M., *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbawn Associates, 1980.
- [3] Jansen, M. G. H., "Estimating the Parameters of a Structural Model for the Latent Traits in Rasch's Model for Speed Tests," *Psychological Measurement*, vol.27(2), pp. 138-151. 2003.

- [4] Yu, M., "Introduction to IRT(12)- computer adaptive test," *Inservice education bulletin*, vol.10(5),. 1993.
- [5] Liu, Z. J., Liang, R. K. & Lin, S. H., "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*, pp. 326–333, Nov. 2001.
- [6] Ho, R. G., "Design a Virtual Item Bank Based on Image Processing Technique,". M, Eng. thesis, Graduate Institute of Information and Computer Education, Taiwan. 2002.
- [7] Battista, M. T., & Clements, D. H., "Student's understanding of three-dimensional rectangular arrays of cubes". *Journal for Research in Mathematics Education*, vol. 27(3), pp. 258-292, 1996.
- [8] Ben-Haim, D., Lappan, G., & Houang, R. T., "Visualizing rectangular solids made of small cubes: Analyzing and effecting student's performance". *Educational Studies in Mathematics*, vol. 16(4), pp. 389-409, 1985.
- [9] Olkun, S., & Knaupp, J. E., *Children's understanding of rectangular solids made of small cubes*, Germany: LAP Lambert Academic Publishing. 2010.
- [10] Chiang, C. T., *Visual-spatial dimensions of cognitive ability*, Taipei: National Changhua University of Education. 1984.