Evaluation for the Test Quality of Dynamic Question Generation by Particle Swarm Optimization for Adaptive Testing

Shu-Chen CHENG, I-Chun PEN & Yu-Chih LIN
Department of Computer Science and Information Engineering
Southern Taiwan University of Science and Technology, Taiwan
kittyc@mail.stust.edu.tw

Abstract: This study discusses the dynamic item-selecting mechanism used in the testing method of Adaptive Testing by Particle Swarm Optimization (PSO). By setting weights of knowledge, we have designed adaptive questions and have evaluated an item's difficulty and discrimination level based on the Item Response Theory (IRT) to see if testees' correct responses match their ability and reach the purpose of adaptive testing. The study designs three experiments to evaluate the quality of the adaptive testing. Namely, they are item discrimination and testee's ability analysis, item exposure rate control and respective test quality.

Keywords: Particle Swarm Optimization (PSO), Adaptive Testing, Item Response Theory (IRT), Item Exposure Rate, Test Quality

1. Introduction

In order to improve the drawbacks of Classical Test Theory (CTT), Lord (1970) proposed the concept of Item Response Theory (IRT)[2][5].Currently, the widely-used models are (1) Single-parameter IRT, (2) Two- parameter IRT, and (3) Three-parameter IRT, respectively. The study proposed the idea to reach the purpose of Computerized Adaptive Testing (CAT) through dynamic multiple-choice questions in order to assess the testee's ability in one specific item in connection with three-parameter IRT and test the correlation among difficulty, pseudo-chance and discrimination of different items and assess item exposure rate and test quality regarding a sample item for the sake of verifying item content and reaching the purpose of "Adaptive learning with student's talent." The objectives of this experiment are to evaluate the quality of adaptive testing implemented by PSO and to explore the correlation between an item's discrimination and difficulty and a testee's ability in regards to three-parameter IRT to see if it matches the weight ratio of knowledge that we established. Furthermore, we use discrimination to study a testee's correct or incorrect response in each item to see if the test is too hard, too simple, as well as to analyze the test's overlap rate and quality for reaching the adaptive testing goal.

2. Literature Review

2.1 Computerized Adaptive Tests

Originally, traditional testing methods were developed on the basis of CTT, in which testees are given the same group of questions. However, this method is not really appropriate for some test types since each testee has different abilities. For a more capable test taker, most questions may be too simple since the taker's ability is not considered in the test with the same questions.

Table 1. Computerized Test Comparison

Type	Computer Based Tests(CBT)	Computerized Adaptive Tests(CAT)
Rationale	Classical Test Theory(CTT)	Item Response Theory(IRT)
Features	1. Same as a traditional paper-based test. 2. Use a computer to assist the test and calculate the score.	1. Test content is generated specifically for an individual testee. 2. Is able to estimate a testee's ability precisely.

CAT is a testing method developed on the basis of IRT. In CAT, test takers are given each item according to their performance in a previous item. When takers finish one item, the system will immediately evaluate their ability and select the next item to be suitable to their ability. In relation to the abovementioned feature of an item level built upon a testee's ability, this may not only lower the item number in an exam, but also precisely evaluate each testee's ability and reach the "Adaptive learning with student's talent"[1][2][4][7][8][9][10][11] idea as listed in Table 1.

2.2 Discrimination Analysis

In brief, discrimination analysis analyzes testees' abilities and responses after they finish a test to discriminate their ability level. A higher discrimination index represents better item discrimination quality while a lower discrimination index may indicate that an item is more difficult or does not match the testee's ability. Usually, discrimination analysis can be divided into a high-score group, P_H and a low-score group P_L. Its equation is listed below:

$$D_i = P_H / P_{HN} - P_L / P_{LN} \quad (1)$$

In this equation, Di is the discrimination of item i; P_H represents how many people in the high-score group have the right answer; P_L represents how many people in the low-score group have the right answer; P_{HN} represents the total number of people in the high-score group; and P_{LN} represents the total number of people in the low-score group.

2.3 Item Response Theory

This is one kind of psychometric theory about one's own response to an individual item, also known as Item response theory. Furthermore, IRT uses equations to describe the possibility of item-response and testee's relative position in a continuity interval. As the name suggests, there are three kinds of parameters in the three-parameter IRT: (1) difficulty, (2) discrimination, and (3) pseudo-chance, whose equation is listed below:

2.3.1 Three-parameter Item Response Theory

Bimbaum (1968)[1] proposed the three-parameter logistic model (3PL), which chiefly deals with the correct possibility as testee n answers item i. As shown in formula (2):

$$P(X_{ni}=1|\theta_n, a_i, b_i, c_i) = c_i + (1-c_i) \times (\exp[a_i(\theta_n - b_i)] / 1 - \exp[a_i(\theta_n - b_i)]) \quad (2)$$

In this formula, X_{ni} is the correct possibility as testee n answers item i; θ_n is the testee's ability; a_i is the discrimination of item i; b_i is the difficulty of item i; and c_i is the pseudo-chance of item i.

2.4 Particle Swarm Optimization

Particle Swarm Optimization (PSO) [3][12][13][13], an algorithm that originated from hunting behavior in bird flocks, is the optimal position-searching technique based on flocking and initially proposed by James Kennedy and Russell Eberhart [3]. Due to its featured simple structure, few parameters, fast convergence and benefits suitable for a dynamic environment.

3. Research Methods

The study aims to provide a test in line with the testee's ability and to reach an adaptive testing objective.

3.1 PSO Computerized Adaptive Dynamic Item-selecting Model

The main purpose of PSO CAT is to provide a test that matches the testee's ability while he/she takes the exam online and decide the difficulty of every following item according to the previous item's result in an attempt to reach computerized adaptive learning goals through the said mechanism. This study selects PSO[3][13][13] as the core algorithm to calculate a great deal of test items in order to find the test that best meets the testee's ability.

3.1.1 PSO Fitness Function

Formula (3) attempts to discover the correlation between the test item and the knowledge block weight that testees themselves define, whose range is $0 \le C_1 \le 1$. A smaller value indicates a higher correlation.

$$C_1 = 1 - \sum_{j=1}^{m} \left(w_j - U_j / T \right) r_j \Big/ q \qquad (3)$$

In this formula, X_j is the exposure control factor whose range is $0 \le X_j \le 1$; U_j is the selected item number in knowledge block j; and T is the estimated item number not selected yet.

Formula (4) balances the frequency of all selected test items. To control the item exposure rate, this study establishes items in a database that was frequently selected to the highest degree and the selected items as evaluation variables. Moreover, to maintain adaptive testing, we select questions that match a testee's ability and related knowledge block as restricted by formula (4) in order to enhance the correct item-selecting rate, which ranges between 0 & 1, $0 \le C2 \le 1$. A smaller value represents less frequency with which the item was already selected.

$$C_2 = n_k \left(1 - C_1\right) / MAX\left(n_1, ..., n_k, ... n_N\right) \qquad (4)$$

In this formula, n_k is the frequency current test item that was already selected and $MAX(n_1,...n_k,...n_N)$ is the highest tested frequency among all current items, $MAX(n_1,...n_k,...n_N) \geq 0$.

Formula (5) is the PSO fitness function, found by adding the previous values of formulas (3) and (4), and is also called the Fitness Value. As the fitness value gets smaller, item difficulty and related knowledge match the testee's current ability better; likewise, this testing suits this testee.

$$Z(I_k) = \left(\sum_{j=1}^{m} |d_k - D| \; r_j \Big/ q \right) + C_1 + C_2 \qquad (5)$$

In this formula, d_k is the item difficulty currently selected; D is the testee's ability level of the current knowledge block; and q represents the knowledge block number relating to the item.

3.1.2 PSO Velocity Function

Velocity function can influence a particle's direction of movement and distance in searching space. Every iteration influences the personal optimal solution and global optimal solution. To avoid falling into a local optimal solution in the search process, this study uses $rand(\,)$ and w as controls and updates a particle's placement as its velocity is acquired by formula (6).

$$v_{i+1} = w \times I_v + k_1 \times rand() \times (I_p - I_k) + k_1 \times rand() \times (I_p - I_k) \qquad (6)$$

In this formula, v_{i+1} is the particle's new velocity after the next iteration; k_1 and k_2 are learning factors; p is the particle's personal optimal solution; g is the global optimal solution; $rand(\,)$ scope is $0 \leq rand(\,) \leq 1$ in order to avoid falling into a local optimal solution; and w is the inertia weight value.

3.2 Item Evaluation Standard

To select an item from its database more evenly and to provide a suitable test, this study also carried out item exposure rates and test quality assessment.

3.2.1 Exposure Rate Evaluation

In an ideal situation, all items in the item back can be selected equally. As an item's exposure rate approaches the average exposure rate, the exposure rate uniformity is better. Formula (7) detects the variance degree of item exposure rate using the Chi-square test.

$$x^2 = \sum_{k=1}^{N} (er_k - \overline{er})^2 \Big/ \overline{er} \qquad (7)$$

In this formula, $er_k = n_k / a$ is the item exposure rate, n_k is the frequency with which the item used is selected, a is the testee number; $\overline{er} = L/N$ is the average item exposure rate, L is test length, and N is test item number; x^2 is the quantitative exposure rate variance of item exposure and average exposure rate. When x^2 becomes smaller, it means that the test tends to be evenly distributed.

3.2.2 Test Quality Evaluation

Formula (8) evaluates the test quality index.

$$TQ = \left\{ DC \Big/ \sum_{u=1}^{S} L_u + \left(\sum_{j=1}^{m} \left| U_j - \frac{w_j}{L} \right| \Big/ 2 \right) + \left(\left(1 - x^2\right) + \left[1 - Max(er)\right] \right) \right\} \Big/ 4 \qquad (8)$$

In this formula, *DM* is the difficulty level coincidence rate used to evaluate if the selected item matches the testee's ability; *DC* is the total item number whose item difficulty matches the testee's ability and knowledge; S is the total test number; L is test length.

3.3 System Structure

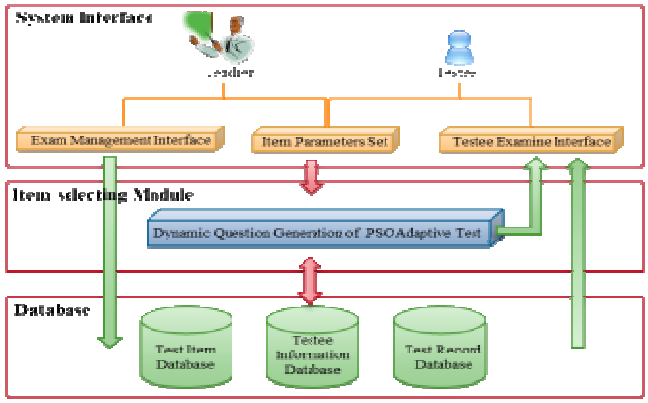

Figure 1. PSO Adaptive Testing Architechture

- System Interface: There are two interfaces, including the teacher setup interface and the testee operation interface, as shown in Figure 1. For the former, the teacher has the ability to add an item into the item database and to design a new test; for the latter, the testee has the ability to establish a weight of knowledge block and item number.
- Item-selecting Module: The study uses computerized adaptive item-selecting system as its core to select suitable test item according to test parameters set in advance and testee ability and display them on testee's testing interface
- Database: The system uses three types of databases. They are test item database, testee information database, test record database respectively.

4. Results

The experiment applied IRT to PSO adaptive testing to estimate the item discrimination and response action of diversified testee abilities, as well as to analyze and compare item exposure rates and test quality. The following are the experimental objectives.

4.1 Item Exposure Rate Control

The experiment took 20 experiments relative to the addition or cancelling of the exposure rate control factor and undertook simulative tests with 25 and 40 items and 100-1000 testees and different-sized databases. The experimental results shown in Figure 2 suggest

that exposure rate addition is superior to cancelling, which slightly improves the item overlap rate problem and proves that exposure rate control factor addition facilitates adaptive dynamic item-selecting construction.

Table 2. 1100-10000-Item Overlap Rate Comparison

Item Overlap Rate					
Item Number	100	500	1000	5000	10000
Add Exposure Rate Factor	0.24	0	0	0	0
Cancel Exposure Rate Factor	0.68	0.52	0.48	0.36	0.44

Table 3. 13000-35000-Item Overlap Rate Comparison

Item Overlap Rate				
Item Number	13000	18000	20000	30000
Add Exposure Rate Factor	0	0	0	0
Cancel Exposure Rate Factor	0.48	0.4	0.28	0.28

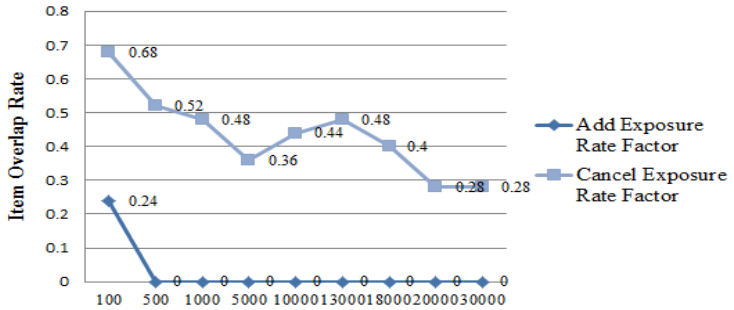

Figure 2. Item Overlap Rate Distribution

4.2 Item Exposure Rate Comparison

This study undertook simulative tests with 25 and 40 items and 500, and 1000 testees and different-sized databases (100, 500, and 1000), respectively. From Tables 4 and 5 below, we know that on the premise that item quantity and test length remain the same, when there are more testees, maximum exposure rate and exposure rate variance tend to decline gradually, meaning that items are not selected evenly. This does not lose control due to an increasing number of people.

Table 4. Exposure Rate Control Comparison of 500 Persons

Number of Testees：500

Length	Number	Max. Exposure Rate	Average Exposure Rate	Variance	Overlap Rate
25	100	46.60%	0.25	3.96	28.80%
25	500	17.80%	0.05	10.96	6.97%
25	100	6.20%	0.025	8.2275	2.83%
40	100	65.60%	0.4	3.341	43.20%
40	500	17.20%	0.08	11.93	10.20%
40	100	9.00%	0.04	12.88	5.10%

Table 5. Exposure Rate Control Comparison of 1000 Persons

Number of Testees：1000

Length	Number	Max. Exposure Rate	Average Exposure Rate	Variance	Overlap Rate
25	100	45.90%	0.25	3.81	28.70%
25	500	17.70%	0.05	10.68	7.04%
25	1000	5.70%	0.025	7.60	2.80%

40	100	66.10%	0.4	3.2847	43.30%
40	500	15.80%	0.08	10.88	10.10%
40	1000	8.80%	0.04	12.167	8.10%

4.3 Test Quality

The experiment used an exam with a 300-500-item database, the test length was 25 items, the number of people was 100, 500, and 1000, and the testee ability default was 0.5. We decided that vocabulary, grammar and reading weights would be 0.4, 0.4, and 0.2, respectively, and the item-select rate is 10, 10, and 5, respectively. The results shown in Figure 3 suggest that increasing items in a database enhance the test quality index; however, in the same testing environment, the results will not be influenced even if the testee number increases. From the experimental results, we also found that the test quality is limited to test length and database ratio. Provided that the minimum test quality tolerance value is 0.8, we can acquire the optimal test quality as the test length and database ratio of 1 to 10.

Table 6. Test Quality Comparison of 300 Persons

Number of Testees：300						
Database Quantity	Number of People	Difficulty	Selected Item Ratio	Max. Exposure Rate	Overlap Rate	Test Quality
300	100	0.77	0.94	0.290	0.1125	0.83
	500	0.77	0.94	0.234	0.108	0.84
	1000	0.77	0.94	0.246	0.113	0.83

Table 7. Test Quality Comparison of 500 Persons

Number of Testees：500						
Database Quantity	Number of People	Difficulty	Selected Item Ratio	Max. Exposure Rate	Overlap Rate	Test Quality
500	100	0.79	0.90	0.200	0.0642	0.86
	500	0.79	0.89	0.178	0.0697	0.86
	1000	0.79	0.88	0.177	0.07	0.85

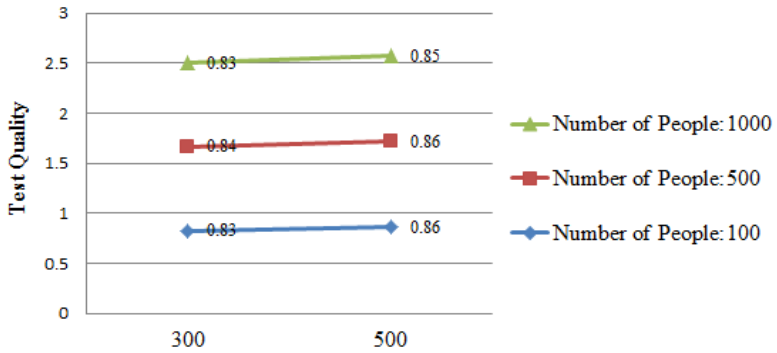

Figure 3. Test Quality Comparisons

5. Conclusion

This study proposed a PSO adaptive item-selecting testing structure to provide testee access in order to select an individual knowledge block level. The experiment's results attempt to analyze item parameters and to analyze discrimination in order to verify whether selected items conform to testee ability. Furthermore, with regard to assessing

overlapped item rate and test quality, experimental results suggest that when the number of people increases, we can get better overlap and exposure rates. For the sake of providing a suitable and excellent test, this study attempts to discuss test quality further; therefore, we used an experiment regarding item difficulty, selected item ratio, exposure rate and overlapped item rate. The results show that test quality is limited to test length and database quantity. The best test quality is obtained in the optimal case when the test length to database quantity ratio is equal to 1:10.

Acknowledgements

This research was partially supported by the National Science Council, Taiwan, ROC, under Contract No.: NSC 99-2511-S-218-007-MY2.

References

[1] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. F. M. L. M. R. Novick, Ed. Reading, *in Statistical theories of mental test scores*, MA:Addison-Wesley, 395-479.

[2] Sympson. J. B., & Hetter, R. D. (1985). Controlling Item Exposure Rates in Computerized Adaptive Testing. *In Proceedings of the 27th annual meeting of the Military Testing Association*, 973-977.

[3] Kennedy, J., & Eberhart, R. C. (1995). Particle Swarm Optimization. *Proceedings of the IEEE International Conference on Neural Networks*, 4, 2113-2122. *Applied Psychological Measurement*, 23(3), 211-222.

[4] Chang, H. H., & Ying, Z. (1999). a-Stratified Multistage Computerized Adaptive Testing.

[5] Lord, F. M. (1997). Practical Applications of Item Characteristic Curve Theory. *Journal of Educational Measurement*, 14, 117-138.

[6] Sun, K. T. (2000). A Genetic Approach to Parallel Test Construction. *International Conference on Computers in Education 2000*, 83-90.

[7] Guzman, E., & Conejo, R. (2005). Self-assessment in a feasible, adaptive web-based testing system. *IEEE Transactions on Education*, 48(4), 688-695.

[8] Huang, T. C., Huang, Y. M., & Cheng, S. C. (2008). Automatic and Interactive e-Learning Auxiliary Material Generation Utilizing Particle Swarm Optimization. *Expert Systems with Applications*, 35(4), 2113-2122.

[9] Shuangyan L., Mike J., & Nathan G. (2008). Incorporating Learning Styles in a Computer-Supported Collaborative Learning Model. *In Proceedings of the 16th International Conference on Computers in Education*, 3-18.

[10] Shu-Chuan, S., Bor-Chen K., & Yu-Lung L. (2008). Building and Applying the Bayesian networks based adaptive test System – Using Rounding & estimating with decimals Unit in the Fifth Grade as A Example. *In Proceedings of the 16th International Conference on Computers in Education*, 117-123.

[11] Huang, Y. M., Lin, Y. T., & Cheng, S. C. (2009). An Adaptive Testing System for Supporting Versatile Educational Assessment. *Computers & Education*, 52(1), 53-67.

[12] Huang, T. C., Cheng, S. C., & Huang, Y. M. (2009). A Blog Article Recommendation Generating Mechanism Using an SBACPSO Algorithm. *Expert Systems with Applications*, 36(7), 10388-10396.

[13] Lin, Y. T., Huang, Y. M., & Cheng, S. C. (2010). An Automatic Group Composition System for Composing Collaborative Learning Groups Using Enhanced Particle Swarm Optimization. *Computers & Education*, 55(4), 1483-1493.

[14] Onjira, S., Tasanawan, S., & Lester, G. (2011). Cognitive Assessment Applying with Item Response Theory. *In Proceedings of the 19th International Conference on Computers in Education*, 339-342.