The Research on Computerized Adaptive Testing

Peng Lu[1,a], **Xiao Cong**[2,b]
[1]Department of Media Technology and Communication, Northeast Dianli University, Jilin, Jilin,132012, P.R.China;
[2]College of Science, Northeast Dianli University, Jilin, Jilin, 132012, P.R.China

E-mail: [a]lup595@nenu.edu.cn; [b]peng.lu2008@gmail.com

Abstract. In recent years, computerized adaptive testing is becoming the focus of the field of modern educational evaluation. In this form of the test, the response relationship between the examinee with the item by IRT modelling, then use the computer to estimates the ability level of the examinees and real-time select item. Computerized adaptive test development process were reviewed in the paper, and discuss the latest research results and pointed out that the current problems and future trends.

1. Introduction

The computerized Adaptive Testing(CAT) is a test form which according to the personalization features of the students. As can be seen from the test development process, as old as the ideas of adaptive testing to students and the test itself, and its prototype is oral in the field of education and psychology in the diagnosis [1], the teachers follow the principle of comparability to students for testing or diagnostic.

The earliest examples of adaptive testing can be traced back to the early 20th century by psychologists Binet. Alfred who developed an intelligence test, the test was conducted precise description to the relationship between the response of student and items. In his view, as long as the same item selection rules which can make a reasonable assessment of all students for a standardized test, and do not need to provide the same items for all students. Therefore, the most important innovation of Benet's test is the intuitive response model[2]. And later, in the study of quantification method, the Psychologist Louis L. Thurstone[3]was used the data sets of the Binet's test. But, at this time in the field of education and psychological tests have been widely used groups test which based on classical test theory, and using observation scores to maintain comparability of scores. This test form to facilitate the implementation, but does not allow any adaptations. For students, the items included in the test are too difficult or too easy and resulting in lower test efficiency.

The development of adaptive testing dependent on two aspects, the most important is the development of the basic theory of test, and it appears theory which called Item Response Theory(IRT). IRT use project characteristic curve (ICC) to explain the response of the probability distribution of the student on the item. The first item response model is two-parameter Normal Ogive Model proposed by Lord [4], and this model as shown in the following equation 1:

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \,, \tag{1}$$

Where, parameter θ represents the ability of student; Discrimination parameter a_i represents the value which is proportional to the slope K at the inflection point of the ICC; difficulty parameter b_i indicates the ability value corresponding to the inflection point at ICC.

Another well-known model is the Rasch model IRT[5], the model includes the ability parameters θ and item difficulty parameter b, and the probability of correct responses of student on an item i is expressed as equation 2:

$$P_i(\theta)=\frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}}, \tag{2}$$

Where, parameter θ represents the ability of student; parameter b_i is difficulty of item i, and has to be estimated sufficiently accurate by check.

In addition to the above item response model, and Birnbaum also proposed Logistic model, Lazarsfeld and Henry proposed potential linear model and potential distance model, Samejima proposed continuous response model, Masters proposed part of the scoring model, etc.. IRT's evolving provides a solid theoretical foundation for adaptive testing.

In addition to the basic theory, on the other hand is the development of computer technology. In the 1970s, Scholars had meaningful attempt on computerized adaptive testing. For example, David Weiss's research on CAT at the University of Minnesota 's done pioneering work [2], and Lord were studied in the computerized estimate the parameters of the items and item selection methods. With the development of computer technology and the continuous improvement of computing power, it has become possible as a testing tool. The large-scale application of adaptive testing appeared in the mid-1990s, it's not only used in the psychological tests, but also widely used in university admission exams, professional certification, military and other fields.

With depth application in practice, the researchers conducted a more in-depth thinking for the basic theory and the actual situation faced in implementation process, it promotes two aspects of continuous development of basic theory of test and computer technology application. We will summarize recent progress and the future development trend of the prospects.

2. The development of basic theory

In recent years, with the application of IRT in the adaptive testing, the researchers have constantly improving for the item response model by adding important parameter, modify the basic assumptions of IRT etc., and in order to be a more scientific evaluation of students, the main contents in detail.

2.1. Item Response Time

Typically, in the test items and students have different response time (RT), and harder items requires more time. And the adaptive algorithm is often providing more time-consuming items for student who has a high level of competency, as a result, the test is not scientific and reasonable. Therefore, automatically recorded the RT of the students in CAT, and the differences of RT into the test results is a new research focus. Representative results in this respect is four-parameter logistic timing model proposed by Tianyou Wang and Bradley A. Hanson [6], as shown in the following equation:

$$P=c_i+(1-c_i)\frac{1}{1+e^{-1.7a_i\left[\theta_m-(\rho_m d_i/t_{mi})-b_i\right]}} \quad , \tag{3}$$

Where, ρ_m is speed parameter of student m, the larger the value, the more time student m needed in resolving item; d_i is speed parameter of item i, the higher the value, the more time required for correct solution item i; t_{mi} is the response time of student m on item i.

The model shows that, when the speed parameters ρ_m and d_i unchanged, increases the probability P of correct responses with the increase of the response time t_{mi} in the item; when then response time t_{mi} in the item unchanged, the greater ρ_m d_i the smaller probability of student m to answer the item i; When t_{mi} tends to infinity, index converges to $a(\theta-b)$. This means that even when time is unlimited it does not guarantee that students will answer item in correct. Therefore, this model is more suitable for proficiency test. With this model can be reasonably estimate of item parameter and the ability for limited testing, and ability to estimate quickly converge to the true ability level. In addition, it can also study whether students can solve items and the accuracy and the best strategy of test.

It should be noted, this model is only a partial description of the test procedure, and a more complete description should include RT distributed on all items in the pool and form a more appropriate statistical model. Therefore, some researchers proposed using a hierarchical framework to improve a convergence rate of the ability estimate at the joint distribution of abilities parameters θ and the speed parameters ρ [7], and proposed the lognormal RT model based on the assuming that the item response time follow a normal distribution of students[8].

Another benefit is that you can use RT abnormalities that may exist in the CAT, such as functional differences, answers deception, and previous knowledge of the items. For these actions, the traditional model-based testing lose effectiveness, and the RT including more information on abnormal behaviour, therefore not affected by this [9]. The RT model with parameter structure allows us to adjust the RT based on the student's actual speed, and check test results their response to the items whether consistent with time-sensitive mode. So, even for those who have experienced cheating, it is impossible to find a regular pattern.

2.2. Multi-grade scoring

So far, most of the operational CAT are based on secondary item response model to deal with objective items[10]. In practice, more attention to students' response on subjective items of test, and many items are multi-grade scores of items, such as calculation items, essay items and so on. The use of multi-grade scoring items can get more information than use two scoring items [11]. Therefore, the use of multi-point scoring model is more reasonable. Most famous multi-point scoring model is Graded Response Mode (GRM) proposed by Samejima[12], the model is described as follows:

Set the full mark of item i is $S(S \geq 1)$, there are $S+1$ number of scoring points, $x = 0,1,2,...,S$; $P_{ix}^{*}(\theta)$ represents the probability that the scores of students who has the ability is θ are not less than x , let scores for all students greater than or equal to x marked "Pass" , and let scores for all students less than x marked "Fail", then $P_{ix}^{*}(\theta)$ has become an item characteristics function to binary scores. And let $P_{i0}^{*}(\theta) = 1$, $P_{i,S+1}^{*}(\theta) = 0$. In this model, uses two stages to obtain the probability of classification score of a student on a certain item.

(1) The first stage, it is compute the cumulative probability to the item i of student who has the ability level of θ , let $P_{ix}^{*}(\theta)$ is 3PLM:

$$P_{ix}^{*}(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{[-Da_i(\theta - b_{ix})]}}, \tag{4}$$

Where, b_{ix} is a threshold parameter which associated with level x of item i, this parameter is subject to the constraint $b_{ix-1} < b_{ix} < b_{ix+1}$. For an item, all response curves share the same discrimination a_i .

(2) The second stage, the response probability obtained on a given level of student. And this probability by subtracting the cumulative probability obtained, as shown below.

$$P_{ix}(\theta) = P_{ix}^{*}(\theta) - P_{ix+1}^{*}(\theta), \tag{5}$$

Figure 1 and Figure 2 show the operator characteristic curve and graded response curve of item has three grades (a=2, b_1 =-1, b_2=0, b_3=1.5, c=0.1).

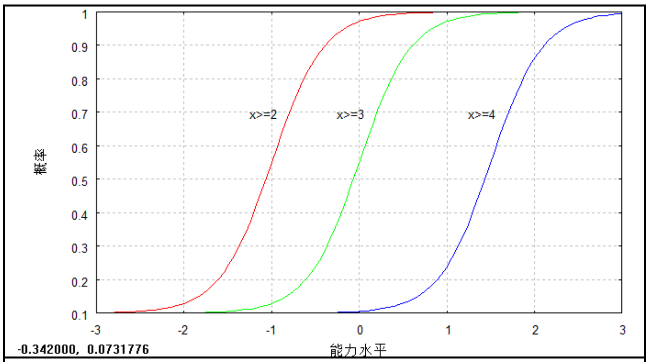

Figure 1. Item characteristic curve

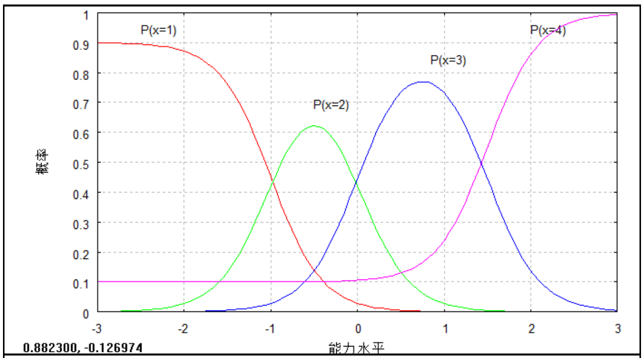

Figure 2. Item Grade response curve

IRT based on multi-level scoring model break out the limit of item response model can be used only two scoring items in the past. Therefore, researchers are increasingly inclined to replace the traditional two score method with multi-level score, also caused the research of more complex scoring models. GRM more applicable to the general items of subjective scoring form and mathematical treatment more convenient. In the coming decades, GRM must have an even broader prospect and good promotional value in applied and theoretical exploration.

2.3. Multidimensional Adaptive Testing

With IRT widely used in practical work, researchers gradually found that the traditional assumption of unidimensional IRT with many actual psychological or educational tests do not match [13]. The multidimensional of test data is consistent with students need to work with a variety of capacities at the completion of a test and few tests measure only a single ability or trait [14]. Therefore, the research on the multidimensional item response theory (MIRT) is very necessary [15], so the researchers' attention gradually shifted to the multidimensional item response models. However, this situation has greatly changed with the development of statistics and computer technology. Therefore, in recent years to study and use of multidimensional item response theory became popular.

Based on the basis of other researchers, Reckase and Mckinley proposed the most practical Logistic multidimensional item response model, and the item response function of this model shown in the following equation [16]:

$$P_{ij}\left(U_{ij}=1\middle|a_i,d_i,c_i,\theta_j\right)=c_i+\left(1-c_i\right)\frac{1}{1+\exp\left(a_i\theta_j+d_i\right)}, \tag{6}$$

Where, $\theta_j=\left(\theta_{j1},\ldots,\theta_{jk},\ldots,\theta_{jm}\right)$ is the parameter of ability vector m of student j , $a_j=\left(a_{j1},\ldots,a_{jk},\ldots,a_{jm}\right)$ is a parameter vector related to the ability to discrimination of items; d_i is a parameter related with difficulty of item, but the meaning is different from the one-dimensional model of the b_i ; and c_i is guess parameter of item. The method of determining all parameters values is through sample tests, collect data and analysed, then determine the value of each parameter.

All items need to show a satisfactory of fit for response model, so the multidimensional items in adaptive test are more dominant role, and should be modelled by multidimensional model and adjust adaptive testing algorithms. If the purpose is for diagnosis, then each dimension should be carefully tested. The change from one-dimensional to multi-dimensional adaptive testing involves an important modification that item selection criteria. For example, the item selection method based on the information function have many parameters during the test, however, due to the information function is replaced by a matrix $p \times p$ (p is the number of item parameters), and the item selection process becomes significantly complex[17]. This not only reflects the estimation accuracy, but also reflects their relevance.

How to reduce multidimensional entity to the one-dimensional standard depends on the purpose of the actual test. For a two-dimensional test, it is should distinguish three different objectives: Firstly, the ability parameters are all the main in two dimensions and should accurately estimate; Secondly, only one main parameter and the other is the interference parameter; Thirdly, merge two parameters to compute their weighted average. For these different goals, it can use the optimize the design principles in statistics for item selection and the rules of assembled item pools to optimize experimental design or sampling procedures.

3. Depth application of computer technology

It is still face many practical problems in the implementation process of CAT, such as in order to maintain the validity of the test, it makes item selection become complex sequences optimization problems with a lot of constrained. And item pool facing high risk and high cost, and how to automatically generate high-quality items in large-scale? The solution to these problems depend on the depth application of computer technology in implementation process of adaptive testing.

3.1. Processing constraints

In the application, the adaptive testing must also meet the safety, comprehensive and many other practical requirements. This methods take test requirements as the constraints of item selection process, and take the adaptive testing as an instance of combinatorial optimization problems for constrained[18].Next, we discuss several major methods.

3.1.1. Weighted deviation modeling method

The weighted deviation modelling (WDM) is proposed by Swanson Stocking [19], it is a heuristic method of item selection. The objective function for item selection is defined the sum of weighted deviation, as shown in the following equation:

$$WDM = \sum_{k=1}^{K} w_k \left(d_{l_k} + d_{u_k} \right) + w_I d_I \quad , \tag{7}$$

Where, K is the number of constraints, $\mathbf{C}$ is the relationship matrix between the items and the constraints, w_k represents the weight of constraint *k*, w_I represents the weight of the amount of information of test; l_k and u_k represent the lower and upper bounds of a constraint *k*; d_{l_k} is the difference compare to lower bound of constraint, d_{u_k} is the excess compare to upper bound of constraint, d_I is the difference compare to information I_t ; e_{l_k} is the excess compare to lower bound of constraint, and e_{u_k} is the difference compare to upper bound of constraint.

The *WDM* method take the item selection constraint problem into a mathematical programming, to select the appropriate item is continuous calculated by the upper and lower bounds. The overall objective function is sum of the difference of all constraints and the difference of information. However, the difference between the various constraints and information are not on the same scale, as exposure constraint is expressed as a percentage, and the content constraint is expressed as the number, therefore, they should not be compared together. In addition, this method requires continuous adjustments to various boundary to achieve the desired result, it is very time consuming [20].

3.1.2. Stratification method

The stratification method is proposed by University of Illinois Professor Hua-Hua Zhang is by Champaign Urbana campus educational psychology at University of Illinois [18]. For exposure of items and content balance problems in CAT, they have proposed STR_A stratification, STR_B stratification and STR_C stratification method.

- STR_A method. The exposure of items which has high distinguish are relatively high, therefore, the ideas of STR_A stratification is stratified item pool based on item discrimination, then the test is divided into several corresponding stages[21]. The studies have shown that this method improves the utilization of low-discrimination items for in the conditions of accuracy.
- STR_B method. Researchers have also proposed STR_B method [22]. The item pool is divided into several blocks according to the item difficulty, then take the items in the various blocks merged into different stratification follow the item discrimination and the item pools generated have similar difficulty distribution in every stratification.
- STR_C method. To solve the problem of the content balance, the researchers proposed STR_C method [23]. The method considers three factors in the process of the stratification of the item pool, such as item discrimination, difficulty and test content. The pool is divided into several groups according to the contents of test; then use STR_B method to get all stratifications in each group.

The stratification approach has made great progress in terms of exposure control and content balance. However, it is need operations such as sort, group, block, stratification and assembly before testing and therefore requires a lot of compute time. It need to reorder and assembly and takes teachers to use too much time in the early preparation, so it is not easy to implement.

3.1.3. Shadow-test method

For the constraint problems faced by item selection process, van der Linden proposed the Shadow-test method [24]. Firstly, it choose a complete test from the item pool and it satisfies all the constraints; then choose the best item from this test based on the initial ability level θ_0 of student; then record student's responses to the item and re-estimate his/her ability level; and according the new ability level θ_1 reassemble test; repeat the above steps until the end of the test.

In the implementation process, the CAT is continuously reassembling Shadow-test, they are only as an intermediate step for item selection, as shown in Figure 3.

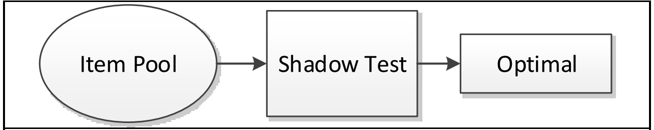

Figure 3. Shadow-test method for adaptive testing

Since each Shadow-test meets all the constraints and use the best item from Shadow-test in the testing process, so adaptive test is also optimal. The method to select the optimal item much faster than an unconstrained adaptive testing. In addition, control the rate of students [25], to ensure that the item exposure in pool and many other practical constraints were studied.

3.1.4. Rotating item pools method

For constraint problem, some researchers from the perspective of the build the item pools proposed rotating item pools (RIP) [26]. In this method, constructed item pools by two stages to handle all of the constraints, as shown in Figure 4.

- The first stage: Assign the items in item pool to a temporary collection. It is calculation and minimize the difference between any two items in item pool by objective function, and take the item pool divided into two similar temporary sets, and so on. Specifically, it can use sequential allocation or simultaneous allocation with heuristic methods.
- The second stage: Assign the items in temporary collection to the child item pools. Depending on the different purpose, it can be constructed of non-overlapping item pools or overlapping item pools. Specific methods can be used random allocation method or mathematical programming methods.

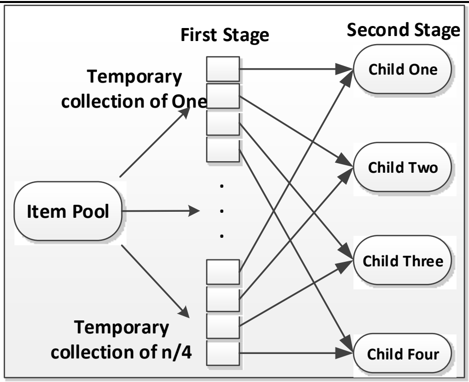

Figure 4. The allocation process of Rotating item pools method

The rotating item pools from design perspective to solve constraint problems. Before the test, take the item pool is divided into several sub-pools which various parameters are almost the same for students to use. And during the test, the system chooses the appropriate item from different sub-pools for students. However, this method also has some problems, such as takes a lot of time in division of the sub-pools before test and required design the number of sub-pools according to students.

3.1.5. The maximum priority index method

The maximum priority index (MPI)[27] take all the constraints variables are weighted by multiplying the value of the maximum amount of information of item and form product PI . Use this maximize product PI as an indicator of measure item instead MI , and the higher the product the greater the priority of the item. The compute of item priority index by the following formula:

$$PI_i = I_i \times \prod_{k=1}^{K} (w_k f_k)^{c_{ik}} \ , \tag{8}$$

Where, K is the total number of all constraints; **C** represents the correlation matrix of items and constraints; I_i is information of item i ; The rest is the product of weight of constraints associated with the item, f_k represents the remaining quota of constraint *k*, w_k represents the weight of constraint *k*. In practice, using a two-stage item selection framework to deal with each flexible constraints which involve upper bounds and lower bounds. It is process the lower bound in the first stage, and process the upper bound in the second stage, but the methods of computing f_k of each phase is different.

Next, the above methods are compared and analysed from the constraint handling, method type, complexity and accuracy affect, and the results shown in Table 1.

Table 1. Comparison of item selection methods

	Constraint handle	Type	Complexity			Accuracy affect
			Prophase	Testing	Maintain	
WDM	many	heuristic method	low	higher	low	yes
STR	less	direct method	high	lower	high	yes
Shadow	more	direct method	low	higher	low	no

RIP	more	direct method	high	lower	high	no
MPI	many	heuristic method	low	higher	low	no

Through the above analysis it can be concluded, the heuristic method can handle multiple constraints such as item exposure, content balance and so on, and the other method is mainly for some constraint problems. However, taking into account the problem of multiple constraints, the calculating of item selection of heuristic method is larger. Overall, the heuristic method is better handling constraints and is the focus of future research.

3.2. Rule-based items generation technology

In traditional tests, it is needs for a specific test to development items and the corresponding pre-tested. If the CAT also follow this thought, so every time a new test will need to replace the entire item pool, this will involve a lot of resources; and if use the same item pool of continuous testing, it will lead to security risk. For this problem, the early solution is to use item exposure control techniques to ensure test security and use of the item pool better. But the researchers then realized, although these methods can improve the utilization rate of project pool project, but it is often difficult to use [28].

Researchers have studied the various types of item automatic generation method; the most representative method is to use an item template. In this method, some of the elements of each item (e.g.: part of stem, correct answer, etc.) have been replaced by the corresponding set. The formation of a number of item family based on the target and use the rules, thus, the entire item pool will be composed by a number item family, shown in Figure 5.

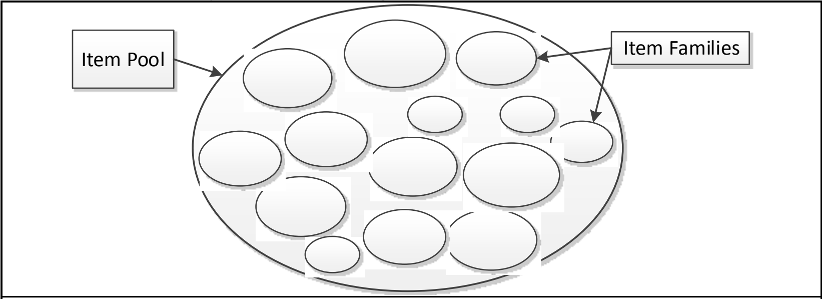

Figure 5. Item family in pool

In practice, the use of two stages of the item selection process for adaptive testing. First of all, according to the current estimate of the student's ability θ , utilizing the difference between the different item families of the item pool to select a best item family; then from this family randomly generated item to students for test, the aim is desired to produce a small difference in the interior of family. Therefore, it need to modify the item response model for a model with a two-stage structure. Set the item family is $f = 1,...,F$, the item in family f is expressed as i_f .Then for the Rasch model, the appropriate two structures is expressed as:

$$P_{i_f} = \frac{e^{\left(\theta - b_{i_f}\right)}}{1 + e^{\left(\theta - b_{i_f}\right)}} \quad , \tag{9}$$

Where, the difficulty parameter b is normal in each family f , and the mean and variance are μ_f and σ_f^2 .So, it can get the difference between the different item families through parameter μ_f and get the difference between the item internal families through parameter σ_f^2 .

Rule-based method for generating an item unresolved problem is the cost of the pre-test. Some researchers believe that the parameters of other items automatically generated from the parameter of an item, so it can save some pre-test activities. In fact, due to the automatically generated items involves more extensive items review, items validation as well as the item parameters vary widely different among item families. In this model, items calibration is turn into item families' calibration.

Therefore, the cost savings of item calibration associated with the number of samples of items and the total number of items generating from the item family.

4. Conclusion and prospect

With the development of item response theory and computer technology also contributed to the development of the CAT and get a lot of important research results. Through computer technology, makes that it is possible to use of sophisticated statistical methods for real-time estimation of ability and to select the best item. Although adaptive testing has made considerable progress, but it has gradually exposed some problems and deficiencies in the years of practice and need to be further studied and improved, it is focused on the following aspects:

(1) More reasonable item response model. For most of the items are multi-grade scores and student will use the ability to use multiple ability to solve items, and response time also reflects the actual situation on the students' ability differences. It is necessary to item response model for a more in-depth research which combine cognitive psychology and statistics, and proposed item response model that meet the characteristics of student and can application in actually.

(2) General constraint handling methods. Like standardized linear tests and in order to maintain the validity, CAT had to meet a wide range of constraints. It should be proposed a universal constraint handling methods to process each constraints using a special algorithm, and can still guarantee a relatively high efficiency and does not affect the user experience. The heuristics algorithm in the field of artificial intelligence computing efficiency, and it can avoid two issues that excessive computing and the feasibility, and provides a viable solution for constraint process.

(3) The organizations of item pool. It can reduce the development costs of the items through using item templates and rule-based automatic generation of high-quality items and maintenance the item parameters by analyse the test record. In addition, using ontology technology to build domain ontology and complete description the details of the relationship between knowledge points and of the contents.

These are the development trend in the next stage of CAT, it also be our further research and exploration. Therefore, through in-depth research on the basic theory and related support technology of CAT, it is can promote the development of the CAT and has important significance for students' assessment of comprehensive quality and the development of education.

Acknowledgments

The work was supported by the Jilin Provincial Education Department "12th Five-Year Plan " scientific and technological research key project(2015239)

References

[1] Linda Crocker, James Algina. Introduction to Classical and Modern Test Theory. Thomson Learning Asia Pte Ltd. (2004).
[2] WimJ. van der Linden. (2008). Some New Developments in Adaptive Testing Technology. Journal of Psychology, 216(1) (2004) 3-11.
[3] Thurstone, L.L. A method of scaling educational and psychological tests. Journal of Educational Psychology, 16 (1925) 433-451.
[4] Peng Lu, Dongdai Zhou, Shaochun Zhou, Xiao Cong. Design and Implementation of Computerized Adaptive Testing System for Multi-Terminals. Modern Educational Technology, 22(6) (2012) 88-92.
[5] Rasch, G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press (1960).
[6] Tianyou Wang, Bradley Allanson . Development and calibration of an Item Response Model that Incorporates Response Time . The annual meeting of the American Educational Research Association in Seattle (2001).

[7] van der Linden, W.J. Using response times for item selection in adaptive testing. Journal of Educational and Behavioral Statistics, 33(1) (2008) 5-20.

[8] van der Linden, W.J. A lognormal model for response times on test items. Journal of Educational and Behavioral Statistics, 31(2) (2006) 181-204.

[9] van der Linden, W.J., & Guo, F.. Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. Psychometrika, 73 (2008) 365-384.

[10] Fen Luo,Shuliang Ding,Xiaoqing Wang. Dynamic and Comprehensive Item Selection Strategies for Computerized Adaptive Testing Based on Graded Response Model. Acta Psychologica Sinica, 44(3) (2012) 400-412.

[11] Ping Chen, Shuliang Ding,Haijing Lin,Jie Zhou. Item Selection Strategies of Computerized Adaptive Testing based on Graded Response Model. Acta Psychologica Sinica, 38(3) (2006) 461-467.

[12] Samejima, F. Estimation of ability using a response pattern of graded scores. Psychometrika, 35(1) (1970) 139-139.

[13] Reckase, M. D. Multidimensional Item Response Theory. Springer-Verlag New York (2009).

[14] Chunhua Kang,Tao Xin. New Development in Test Theory: Multidimensional Item Response Theory. Advances in Psychological Science, 18(3) (2010) 530-536.

[15] Dongbo Tu, Yan Cai, Haiqi Dai, Shuliang Ding. Parameters Estimation of MIRT Model and Its Application in Psychological Tests. Acta Psychologica Sinica, 43(11) (2011) 1329-1340.

[16] Reckase, M. D., & McKinley, R. L. Some Latent Trait Theory in a Multidimensional Latent Space. Iowa City, IA: American College Service (1982).

[17] Mulder, J., & van der Linden, W.J. Multidimensional adaptive testing with optimal design criteria for item selection . Psychometrika, 74 (2009) 273-296.

[18] Peng Lu, Dongdai Zhou, Shaochun Zhou, Xiao Cong. Research on Item Selection Method For CAT Based On Simulated Annealing . Computer Applications and Software, 29(10) (2012) 175-179.

[19] Swanson, I.., & Stocking, M. I. A model and heuristic for solving very large item selection problems . Applied Psychological Measurement, 17 (1993) 151-166.

[20] Leung C. K., Chang, H. H. and Hau, K. T. Computerized adaptive testing: A mixture item selection approach for constrained situations. British Journal of Mathematical & Statistical Psychology, 58 (2005) 239-257.

[21] Hau, K. T., & Chang, H. Item selection in computerized adaptive testing: Should more discriminating items be used first?. Journal of Educational Measurement, 38 (2001) 249-266.

[22] Chang, H., Qian, J., & Ying, Z. a-stratified multisage CAT with b-blocking. Applied Psychological Measurement, 25 (2001) 333-341.

[23] Yi, Q., & Chang, H. a-stratified CAT design with content blocking. British Journal of Mathematical and Statistical Psychology, 56 (2003) 359-378.

[24] van der Linden, W.J. Linear models for optimal test design. New York: Springer-Verlag (2005).

[25] van der Linden, W.J., Breithaupt, K., Chuah, S.C., & Zhang, Y. Detecting differential speed in multistage testing. Journal of Educational Measurement, 44 (2007) 117-130.

[26] Ariel, A., Veldkamp, B. P., & van der Linden, W. J. Constructing rotating item pools for constrained adaptive testing. Journal of Educational Measurement, 41(2004) 345-360.

[27] Cheng, Y., & Chang, H. The maximum priority index method for severely constrained item selection in computerized adaptive testing. British Journal of Mathematical and Statistical Psychology, 62 (2009) 369-383.

[28] Chang, H., & Ying, Z. a-Stratified multistage computerized adaptive testing. Applied Psychological Measurement, 23 (1999) 211-222.