An Introduction to the Computerized Adaptive Testing

TIAN Jian-quan, MIAO Dan-min, ZHU Xia, GONG Jing-jing
(School of Aerospace, Fourth Military Medical University, Xi'an Shaanxi 710032, China)

Abstract: The computerized adaptive testing (CAT) has unsurpassable advantages over the traditional testing. It has become the mainstream in large scale examination in modern society. This paper gives a brief introduction to CAT including differences between traditional testing and CAT, the principals of CAT works, Psychometric theory and computer algorithms of CAT, the advantages and cautions of CAT. In the end, the development of CAT in China is reviewed.

Key words: computerized adaptive testing; item response theory; psychological measurement; educational measurement

The computerized adaptive testing (CAT) has become increasingly common in educational assessment in the United States, especially in large-scale testing programs including the Graduate Record Examination (GRE), the Graduate Management Admission Test (GMAT), and the Armed Services Vocational Aptitude Battery (ASVAB). In China, research about item response theory (IRT) and CAT has been conducted more than 30 years, but the application of CAT is just in the course of exploration and development. It's necessary to give a simple introduction about CAT.

1. Differences between Traditional Testing and Adaptive Testing

Table 1 Differences between traditional testing and adaptive testing

	Traditional test	Adaptive test
Composition of test	Each candidate takes an identical test.	Each candidate takes a different test.
Difficulty of test	Aimed at the average candidate.	Aimed at the individual candidate.
Test length	Identical for each candidate. Comparatively longer tests.	Different for each candidate. Comparatively shorter tests.
Test moment	Fixed moment at which all students are tested.	Any moment suitable to the student.
Test organization	Much time needed.	Little time needed.
Immediate results	No	Yes

From table 1 we can see that an adaptive test is a test of which the items are taken from a large item bank

TIAN Jian-quan, Ph.D. candidate, lecturer of Department of Psychology, School of Aerospace, Fourth Military Medical University; research field: applied psychology.
MIAO Dan-min, Ph.D., professor of Department of Psychology, School of Aerospace, Fourth Military Medical University; research field: psychological measurement.
ZHU Xia, Ph.D., associate professor of Department of Psychology, School of Aerospace, Fourth Military Medical University; research field: psychological measurement.
GONG Jing-jing, Ph.D. candidate, lecturer of Department of Psychology, School of Aerospace, Fourth Military Medical University; research field: cognitive psychology.

during the administration of the test. The selection of an item depends on the candidate's proficiency estimated at that moment. If the proficiency is estimated to be relatively high, the candidate is presented with a relatively difficult item. If the proficiency is estimated to be relatively low, the candidate has to answer a relatively simple item. The combination of this new principle of testing and the use of a computer for the administration of a test show considerable differences between traditional pencil-and-paper tests and adaptive tests. We can have a look at some of these differences now.

2. The Principles of CAT Works

2.1 Building an item bank

A necessary pre-requisite to computer-adaptive testing is an item bank (Wright and Bell, 1984). An item bank is an accumulation of test items. There is the text of the item, details of correct and incorrect response to it, and its current difficulty estimate. If the item has a rating scale or internal scoring structure, that is also included. There may also be indicators of item content area, instructional grade level and the like. Usually it is also include details of the course, in which the items are developed, used and recalibrated.

Initially CAT item banks usually contain items given under conventional paper-and-pencil conditions. For any particular test in that format, every item has been given to every test-taker. This enables at least a p-value (percent of success on the item for the sample) for each item to be computed. An initial estimate of the logic difficulty of an item within a test form then becomes log (100-pvalue/pvalue). Available Rasch software, e.g., BIGSTEPS (Linacre & Wright, 1988) enables production of better initial item difficulty estimates. Test equating procedures (Wright & Stone, 1979) enable the difficulties of all items to be estimated within one common frame of reference. These items can then be entered into an item bank, and CAT administration begins fairly quickly. Studies have indicated that most paper-and-pencil items maintain their difficulty level when transferred to CAT. Exceptions are items with idiosyncratic presentation requirements. For instance, some figures and graphical plots are easier to think about (and make annotations on), when they are presented horizontally on a paper-and-pencil test, than when they are presented vertically on a CAT computer screen.

When an item bank is to be constructed out of newly composed items, difficulty levels must be newly assigned other than directly from p-values or quantitative item analysis. Stratifying or ordering items by difficulty has two aspects. First, there is ordering based on the theoretical construct. Experts in a field generally know what topic areas should be harder than others for those at any stage of development. This enables an ordering of items by topic area difficulty. In addition, inspection of individual items gives indications of their relative difficulty. Consequently, a fairly robust stratifying of items by expert-perceived difficulty can often be accomplished. There are situations, however, when there is no clear construct-based ordering. A multiple-choice question (MCQ) may be written with its incorrect options, i.e., distracters, so close to, or far from, the correct answer as to render the item much harder, or much easier, than it should be according to its construct level. Second, there is ordering based on empirical performance of a previous sample of test candidates. For brand new items, this does not exist, but it is often possible to identify similar pre-existing items. Then the difficult levels of these items can be used.

For larger scale testing, testing agencies often enter into CAT with an accumulation of items of uncertain quality and dimensionality. An advantage of the CAT approach is that changes of the item bank can be made at any point in test administration. There is no need to wait for the last test-taker to complete the test before item analysis can begin. Item analysis should be conducted concurrently with test administration. This validates not

only that item selection and ability measurement are functioning correctly, but also that the items themselves are functioning at their specified difficulty levels. Recent experience with the CAT version of the GRE, Graduate Record Examination (Smith, 1999) is a reminder that there must be a continuous program of quality control and test validation for CAT, just as much as for other testing methods.

A virtue of CAT is that the new items can be introduced into the bank easily. Initially, new items can be administered inconspicuously along with pre-existing items, but not used for test-taker ability estimation. Instead, the test-takers' responses are used to verify the item is functioning as specified and to ascertain the item's precise difficulty. Then the item can be made part of the regular bank.

When an item is revised it becomes a new item. Revision must change some aspect of the item. So it must impact the item's difficulty, or some other aspect of the item's functioning. Consequently a revised item must be regarded as a new item, and its difficulty re-estimated accordingly.

CAT testing is often done at remote locations. Under these circumstances, the item bank, even if encrypted and otherwise secured, should not be transported in its entirety to all locations. Instead, different locations should be sent different, overlapping, sections of the item bank. This has several benefits. First, test security is improved because the theft of one test package does not compromise the entire bank. Secondly, item exposure is limited. Any item can only be seen by a fraction of the candidates, at most. Thirdly, the chances of a large number of test-takers experiencing identical tests are diminished overall. Fourthly, if problems are discovered during test administration, afterwards, only a fraction of the CAT administrations is likely to be affected. Fifthly, the overlap is introduced so that item difficulties at different sites can be compared and equated, thus insuring a fair evaluation of performance for all test-takers.

2.2 The administration of CAT

There are two types of item, one is dichotomous items, to which the responses are scored by "right" or "wrong", corresponding scores are "1" or "0". The other is polychromous items, its score rules are complicated, and so we use dichotomous items as examples to introduce the administration of CAT.

Imagine that an item bank has been constructed of dichotomous items, e.g., of multiple-choice questions (MCQs). Every item has a difficulty expressed as a linear measure along the latent variable of the construct. For ease of explanation, let us consider an arithmetic test. The latent variable of arithmetic is conceptually infinitely long, but only a section of this range is relevant to the test and is addressed by items in the bank. Let us number this section from 0 to 100 in equal-interval units. So, every item in the bank has a difficulty in the range 0 to 100. Suppose that 2+2=4 has a difficulty of 5 units. Children for whom 2+2=4 is easy have ability higher than 5 units. Children for whom 2+2=4 is too difficult to accomplish correctly have ability below 5 units. Children with a 50% chance of correctly computing that 2+2=4 have an estimated ability of 5 units, the difficulty of the item. This item is said to be "targeted on" those children.

Here is how a CAT administration could proceed. The child is seated in front of the computer screen. Two or three practice items are administered to the child in the presence of a teacher to ensure that the child knows how to operate the computer correctly. Then the teacher keys in to the computer an estimated starting ability level for the child, or, the computer selects one by itself.

Choice of the first question is not critical to measurement, but it may be critical to the psychological state of the candidate. Administer an item that is much too hard, and the candidate may immediately fall into despair, and not even attempt to do well. This is particularly the case if the candidate already suffers anxiety about the test. Administer an item that is much too easy, and the candidate may not take the test seriously and so make careless

mistakes. Gershon (1992) suggests that the first item, and perhaps all items, should be a little on the easy side, giving the candidate a feeling of accomplishment, but in a situation of challenge.

If there is a criterion pass-fail level, then a good starting item has difficulty slightly below that. Then candidates with ability around the pass-fail level are likely to pass, and to know that they passed, that first item and so be encouraged to keep trying.

For example, suppose that the first item to be administered is of difficulty 30 units, but that the child has ability 50 units. The child will probably pass that first item. Let's imagine what happens (see Figure 1). The computer now selects a more difficult item, one of 40 units. The child passes again. The computer selects a more difficult item, one of 50 units. Now the child and the item are evenly matched. The child has a 50% chance of success. Suppose the child fails, the computer administers a slightly easier item than 50 units, but harder than the previous success at 40 units. A 45 unit item is administered. The child passes. The computer administers a harder item at 48 units. The child passes again. In view of the child's success on items between 40 and 48 units, there is now evidence that the child's failure at 50 may be unlucky.

The computer administers an item of difficulty 52. This item is only slightly too hard for the child. The child has almost a 50% chance of success. In this case, the child succeeds. The computer administers an item of difficulty 54 units. The child fails. The computer administers an item of 51 units. The child fails. The computer administers an item of 49 units. The child succeeds.

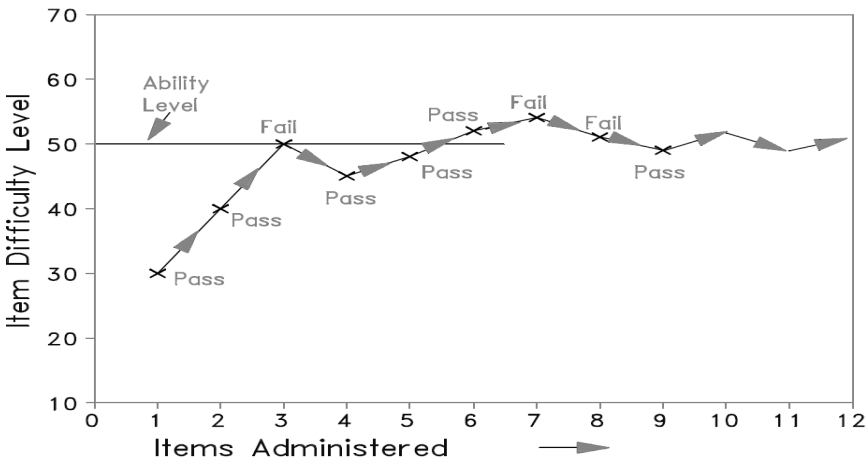

Figure 1 Dichotomous CAT test administration

This process continues. The computer program becomes more and more certain that the child's ability level is close to 50 units. The more items that are administered, the more precise this ability estimate can be. The computer program contains various criteria, "stopping rules", for ending the test administration. When one of these is satisfied, the test stops. Then the results of the test are reported (or stored) by computer. The candidate is dismissed and the testing of the next candidate begins.

There are often other factors that also affect item selection. For instance, if a test address a number of topic areas, then content coverage may require that the test include items be selected from specific subsets of items. Since there may be no item in the subset near the candidate's ability level, some content-specific items may be noticeably easier or harder than the other items. It may also be necessary to guard against "holes" in the candidate's knowledge or ability or to identify areas of greater strength or "special knowledge". The occasional

administration of an out-of-level item will help to detect these. This information can be reported diagnostically for each candidate, and also used to assist in pass-fail decisions for marginal performances.

The dichotomous test is not one of knowledge, ability or aptitude, but of attitude, opinion or health status, then CAT administration follows the same plan as above. The difference is that the test developer must decide in which direction the variable is oriented. Is the answer to be scored as "right" or "correct" to be the answer that indicates "health" or "sickness"? Hire "right" or "correct" is to be interpreted to be "indicating more of the variable as we have defined the direction of health or sickness." The direction of scoring will make no difference to the reported results, but it is essential in ensuring that all items are scored consistently in the same direction. If the test is to screen individuals to see if they are in danger of a certain disease, then the items are scored in a direction such that more danger implies a higher score. Thus the "correct" answer is the one indicating the greater danger.

3. Psychometric Theory and Computer Algorithms of CAT

3.1 Choice of the measurement model

An essential concept underlying almost all ability or attitude testing is that the abilities or attitudes can be ranked along one dimension. This is what implied when it is reported that one candidate "scored higher" than another on a certain test. If scores on a test rank candidates in their order of performance on the test, then the test is being used as though it ranks candidates along a one-dimensional variable.

Of course, no test is exactly one-dimensional. But if candidates are to be ranked either relative to each other, or relative to some criterion levels of performance (pass-fail points), then some useful approximation to unidimensionality must be achieved.

Unidimensionality facilitates CAT, because it supports the denotation of items as harder and easier, and test-takers as more and less able, regardless of which items are compared with which test-takers. Multidimensionality confounds the CAT process because it introduces ambiguity about what "correct" and "incorrect" answers imply. Consider a math "word problem" in which the literacy level required to understand the question is on a par with the numerical level required to answer the question correctly. Does a wrong answer mean low literacy, low numeracy or both? Other questions must be asked to resolve this ambiguity, implying the multidimensional test is really two one-dimensional tests intertwined. Clearly, if the word problems are intended to be a math test, and not a reading test, the wording of the problems must be chosen to reduce the required literacy level well below that of the target numeracy level of the test. Nevertheless, investigations into CAT with multidimensionality are conducted (Vander Linden, 1999).

Since it can be demonstrated that the measurement model necessary and sufficient to construct a one-dimensional variable is the Rasch model (e.g., Wright, 1988), the discussion of CAT algorithms will focus on that psychometric model. Even when other psychometric models are chosen initially because of the nature of pre-existing item banks, the constraints on item development in a CAT environment are such that a Rasch model must then be adopted. This is because test-takers are rarely administered items sufficiently off-target to clearly signal differing item discriminations, lower asymptotes (guessing) or higher asymptotes (carelessness). Similarly, it is no longer reasonable to assert that any particular item was exposed to a normal (or other specified) distribution of test-takers. Consequently, under CAT conditions, the estimation of the difficulty of new items is reduced to a matter of maintaining consistent stochastic ordering between the new and the existing items in the

bank. The psychometric model necessary to establish and maintain consistent stochastic ordering is the Rasch model (Roskam & Jansen, 1984).

A concern can arise here that both test-takers and items are being located along the same ability scale. How can the items be placed on an ability scale? At a semantic level, Andrich (1990) argues that the written test items are merely surrogate, standardized examiners, and the struggle for supremacy between test-taker and item is really a struggle between two protagonists, the test-taker and the examiner. At a mathematical level, items are placed along the ability metric at the points at which those test-takers have an expectation of 50% success on those items.

This relationship between test-takers and items is expressed by the dichotomous Rasch model (Rasch, 1960/1992):

$$\log\left(\frac{P_{n/1}}{P_{n/0}}\right) = B_n - D_i$$

Where P_{ni1} is the probability that test-taker n succeeds on item *i*, and P_{ni0} is the probability of failure. The natural unit of the interval scale constructed by this model is termed the logit (log-odds unit). The logit distance along the one-dimensional measurement scale between a test-taker expected to have 50% success on an item, (i.e., at the person at same position along the scale as the item), and a test-taker expected to have 75% success on that same item is log (75%/25%)=1.1 logits.

From the simple, response-level Rasch model, a plethora of CAT algorithms have been developed.

3.2 The design of the algorithm

In essence, the CAT procedure is very simple and obvious. A test-taker is estimated (or guessed) to have a certain ability. An item of the equivalent level of difficulty is asked. If the test-taker succeeds on the item, the ability estimate is raised. If the test-taker fails in the item, the ability estimate is lowered. Another item is asked, targeted on the revised ability estimate. And the process repeats. Different estimation algorithms revise the ability estimate by different amounts, but it has been found to be counter-productive to change the ability estimate by more than 1 logit at a time. Each change in the ability estimate is smaller, until the estimate is hardly changing at all. This provides the final ability estimate.

3.3 Stopping rules

The decision as to when to stop a CAT test is the most crucial element. If the test is too short, then the ability estimate may be inaccurate. If the test is too long, then time and resources are wasted, and the items exposed unnecessarily. The test-taker also may tire, and drop in performance level, leading to invalid test results.

The CAT test stops when:

(1) The item bank is exhausted.

This occurs, generally with small item banks, when every item has been administered to the test-taker.

(2) The maximum test length is reached.

There is a pre-set maximum number of items that are allowed to be administered to the test-taker. This is usually the same number of items as on the equivalent paper-and-pencil test.

(3) The ability measure is estimated with sufficient precision.

Each response provides more statistical information about the ability measure, increasing its precision by decreasing its standard error of measurement. When the measure is precise enough, testing stops. A typical standard error is 0.2 logits.

(4) The ability measure is far enough away from the pass-fail criterion.

For CAT tests evaluating test-takers against a pass-fail criterion level, the test can stop once the pass-fail decision is statistically certain. This can occur when the ability estimate is at least two S.E.'s away from the criterion level, or when there are not sufficient items left in the test for the candidate to change the current pass-fail decision.

(5) The test-taker is exhibiting off-test behavior.

The CAT program can detect response sets (irrelevant choice of the same response option or response option pattern), responding too quickly and responding too slowly. The test-taker can be instructed to call the test supervisor for a final decision as to whether to stop or postpone the test.

The CAT test cannot stop before:

(1) A minimum number of items have been given.

In many situations, test-takers will not feel that they have been accurately measured unless they have answered at least 10 or 20 items, regardless of what their performances have been. They will argue, "I just had a run of bad luck at the start of the test, if only you had asked me more questions, my results would have been quite different!"

(2) Every test topic area has been covered.

Tests frequently address more than one topic area. For instance, in arithmetic, the topic areas are addition, subtraction, multiplication and division. The test-taker must be administered items in each of these four areas before the test is allowed to stop.

(3) Sufficient items have been administered to maintain test validity under challenge or review.

This can be a critical issue for high-stakes testing. Imagine that the test stops as soon as a pass or fail decision can be made on statistical grounds. Then those who are clearly expert or incompetent will get short tests, marginal test-takers will get longer tests. Those who receive short tests will know they have passed or failed. Those who failed will claim that they would have passed, if only they had been asked the questions they know. Accordingly it is prudent to give them the same length test as the marginal test-takers. The experts, on the other hand, will also take a shorter test, and so they will know they have passed. This will have two negative implications. Everyone still being tested will know that they have not yet passed, and may be failing. Further, if on review it is discovered there is a flaw in the testing procedure, it is no longer feasible to go back and tell the supposed experts that they failed or must take the test again. They will complain, "Why didn't you give me more items, so that I could demonstrate my competence and that I should pass, regardless of what flaws are later discovered in the test."

4. The Advantages of CAT

Many of the advantages of CAT have been indicated in the preceding discussion, here are the advantages identified by Rudner (1998).

(1) In general, computerized testing greatly increases the flexibility of test management. (e.g. Weiss & Kingsbury, 1984)

(2) Tests are given "on demand" and scores are available immediately.

(3) Neither answer sheets nor trained test administrators are needed. Test administrator differences are eliminated as a factor in measurement error.

However, supervision is still needed, and the environment in which CAT is conducted can definitely affect test results.

(4) Tests are individually paced so that examinee does not have to wait for others to finish before going on to the next section. Self-paced administration also offers extra time for examinees who need it, potentially reducing one source of test anxiety.

(5) Test security may be increased because hard copy test booklets are never compromised.

Further, if no two people take the same test, parroting answers or copying from someone else is pointless.

(6) Computerized testing offers a number of options for timing and formatting. Therefore it has the potential to accommodate a wider range of item types.

These can include moving images, sounds, and items that change their appearance based on responses to previous items.

(7) Significantly less time is needed to administer CATs than fixed-item tests since fewer items are needed to achieve acceptable accuracy. CATs can reduce testing time by more than 50% while maintaining the same level of reliability.

(8) Shorter testing times also reduce fatigue, a factor that can significantly affect an examinee's test results.

(9) CATs can provide accurate scores (measures) over a wide range of abilities while traditional tests are usually most accurate for average examinees.

A CAT test differs profoundly from a paper-and-pencil (P&P) test. The primary advantage of a CAT to test developers and administrators is its promise of efficient testing. In theory, examinee testing times can be dramatically reduced while maintaining the quality of measurement provided by conventional (i.e., fixed-item) tests. This advantage is particularly attractive to testing programs that have traditionally required lengthy tests. In such testing contexts, the potential problem of examinee fatigue and, consequently, diminished effort can be alleviated by use of a CAT. Virtually all operational CATs use measurement methods based on item response theory (IRT) (Lord & Novick, 1968) to select test items to administer and to estimate examinee proficiency. The invariance principle of IRT allows one to administer different sets of items drawn from an item pool to different examinees, yet estimate their relative levels of proficiency on a common scale of measurement. A CAT's efficiency is realized through the targeting of item difficulty to examinee proficiency. IRT principles suggest that items targeted in this manner provide maximal information in the estimation of examinee proficiency.

5. Cautions with CAT

Here are the limitations to CAT identified in Rudner (1998).

(1) CATs are not applicable for all subjects and skills. Most CATs are based on an item response theory model, yet item response theory is not applicable to all skills and item types.

This is true. Similar limitations apply to paper-and-pencil tests.

(2) Hardware limitations may restrict the types of items that can be administered by computer. Items involving detailed art work and graphs or extensive reading passages, for example, may be hard to present.

Advances in computer technology and better item presentation are eliminating many of these concerns.

(3) CATs require careful item calibration. The item parameters used in a paper and pencil testing may not hold with a computer adaptive test.

As Wright and Douglas (1975) and other studies show, there is no exact item calibration. Neither is there a need for the estimated difficulty of CAT items to exactly match the paper-and-pencil estimated difficulties. In fact, because of the more relevant sample, the CAT item difficulties should be more believable.

(4) CATs are only manageable if a facility has enough computers for a large number of examinees and the examinees are at least partially computer-literate. This can be a big limitation.

The extent of this limitation depends on the reason for the test and the characteristics of the test-takers. Classroom level CAT can be done on one computer by one child at a time. Low stakes tests can be done via the Internet. Rudner is here referring to large-scale tests such as the SAT and ACT. These are already under more powerful attack for other reasons.

(5) The test administration procedures are different. This may cause problems for some examinees.

As computers become more pervasive, it may be the paper-and-pencil tests, with their bubble sheets, that are seen as problematic.

(6) With each examinee receiving a different set of questions, there can be perceived inequities.

This is why it is essential that every test-taker be administered enough items to insure that their final ability estimate is unassailably reasonable.

(7) Examinees are not usually permitted to go back and change answers.

Improved item selection and ability estimation algorithms now allow test-takers to review and change previous responses.

(8) If changing responses is permitted, a clever examinee could intentionally miss initial questions. The CAT program would then assume low ability and select a series of easy questions. The examinee could then go back and change the answers, getting them all right. The result could be 100% correct answers which would result in the examinee's estimated ability being the highest ability level.

This has been investigated both in practice and statistically, and found to be a wild gamble. It based on the incorrect notion that a perfect score on an easy test will result in an ability estimate at the highest level. In fact, with effective CAT algorithms, such as those of Halkitis or UCAT, it will not.

Gershon and Bergstrom (1995) considered this strategy under the best possible conditions for the potential cheater: A CAT test which allows an examinee to review and change any responses. This type of examinee-friendly CAT is already used in high-stakes tests and will rapidly spread, once CAT fairness becomes a priority.

6. The Development of CAT in China

In China, the research about item response theory and CAT has been conducted for more than 30 years. Researchers had been familiar with topics including development of item bank, item calibration and test equating. Software such as ANOTE has been developed to estimate item parameter and to resolve test equating. Further more, some computerized adaptive tests were developed, such as the computerized adaptive test of language proficiency for middle school students and computerized adaptive test of mathematic ability proficiency for primary school students (QI Shu-qing, 2003). Large scale examination such as HSK has been conducted country wide. However, most research about CAT is theoretical. In order to meet the practical need, much effort should be made.

References:

[1] Wright, B.D. & Bell, S.R. Item Banks: What, Why, How. *Journal of Educational Measurement,* 1984(21): 331-345.

[2] Linacre, J.M. & Wright, B.D. *BIGSTEPS: Rasch Measurement Computer Program*. Chicago: Mesa Press, 1988.

[3] Wright, B.D. & Stone, M.H. *Best Test Design*. Chicago: Mesa Press, 1979.

[4] Gershon, R.C. Test Anxiety and Item Order: New Concerns for Item Response Theory. Chapter 11//M Wilson. ed. *Objective Measurement: Theory into Practice*, Norwood, NJ.: Ablex. 1992, 1.

[5] Vander Linden, W.J. Multidimensional Adaptive Testing with a Minimum Error-variance Criterion. *Journal of Educational and Behavioral Statistics*, 1999 (24):4, 398-412.

[6] Wright, B.D. Rasch model from Campbell Concatenation. *Rasch Measurement Transactions*, 1988, 2(1): 16.

[7] Roskam, E.E. & Jansen, P.G.W. A New Derivation of the Rasch Model//E. Degreef & J. van Bruggenhaut, eds. *Trends in Mathematical Psychology*. Amsterdam: North-Holland, 1984: 293-307.

[8] Andrich, D.A. The Ability of an Item. *Rasch Measurement Transactions*, 1990, 4(2): 101.

[9] Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen and Chicago: Mesa Press, 1960/1992.

[10] Rudner, L. *An On-line, Interactive, Computer Adaptive Testing Mini-Tutorial*. ERIC Clearinghouse on Assessment and Evaluation, 1998.

[11] Lord, F. M. & Novick M. R. *Statistical Theory of Mental Test Scores*. Reading, MA: Addison-Wesley, 1968.

[12] Wright, B.D. & Douglas, G. *Best Test Design and Self-tailored Testing*. MESA Memorandum No. 19. Department of Education, Univ. of Chicago, 1975.

[13] Weiss, D.J. & Kingsbury, G.G. Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*, 1984, 21(4): 361-375.

[14] E. Degreef & J. van Bruggenhaut, eds. *Trends in Mathematical Psychology*. Amsterdam: North-Holland, 1984.

[15] Gershon, R.C. & Bergstrom, B. *Does Cheating on CAT Pay: NOT*. Paper presented at The Annual Meeting of the American Educational Research Association. Chicago, 1995.

[16] QI Shu-qing. *The Application of Modern Measurement Theory in Examination*. Wuhan: Huazhong Normal University Press, 2003: 350-369.

(Edited by CHEN Jing and ZHANG Dong-ling)

(continued from Page 68)

References:

[1] Ellis, R. *The study of second language acquisition*. Oxford: Oxford University Press, 1994.

[2] Tricial Hedge. *Teaching and Learning In the Language Classroom*. Shanghai: Foreign Language Education Press, 2002.

[3] Jeremy Harmer. *The practice of English Language Teaching*. Longman, 2003.

[4] Andrew D. Cohen 2000. *Strategies in Learning and Using a Second Language* Foreign language Teaching and Research Press.

[5] Michael J. Wallace. *Action Research for Language Teachers*. Cambridge: Cambridge University Press, 1998.

[6] Monk, G. Stephen. *Student Engagement and Teacher Power in Large Classes: Learning in Groups*. In Clark Bouton and Russell Y. Garth. (Ed.), New Directions for Teaching and Leaning. No.14. San Francisco: Jossey-Bass.

[7] Nolasco, R.& Arthur, L. *Teaching Large Class*. Macmillan, 1988.

[8] Shi Liangfang. *Theory of Learning*. Beijing: People's Education Press, 2001.

(Edited by CHEN Jing and ZHANG Dong-ling)