A Historical Review of the Theoretical and Technological Development of Computerized Adaptive Testing

Jian Xiaozhu

2026-08-23

ChinaRxiv

Abstract

Computerized adaptive testing (CAT) originated from the Binet-Simon Intelligence Test. In the early 1970s, CAT technology, grounded in item response theory as its measurement framework, achieved breakthrough technological development. Based on the forms of CAT, this paper divides the development of CAT into five stages: the embryonic stage of adaptive testing, the early stage, the breakthrough stage, the initial development stage of CAT, and the current development stage. This paper discusses the forms of test development, main characteristics, and practical applications of these five stages in sequence, and further examines and prospects hot research topics concerning the application of artificial intelligence (AI) technology to CAT.

Full Text

A Historical Review of the Theoretical and Technological Development of Computerized Adaptive Testing

Jian Xiaozhu*

(School of Public Policy and Administration, Nanchang University, Nanchang, Jiangxi, 330036)

Abstract: Computerized adaptive testing (CAT) originated with Binet-Simon intelligence testing. In the early 1970s, CAT testing technology achieved a breakthrough in technique by taking item response theory as its measurement-theoretical basis. According to the forms of CAT testing, this paper divides the development process of CAT into five stages: the embryonic stage of adaptive testing, the early stage, the breakthrough stage, the initial development stage of CAT, and the current development stage. This paper discusses, in sequence, the forms of test development, main characteristics, and practical applications in these five stages, and reviews and looks ahead to hot research topics concerning the application of artificial intelligence (AI) technology to CAT.

Keywords: adaptive testing; computerized adaptive testing; item response theory; measurement theory

1 Introduction

Computerized adaptive testing (CAT) selects items of appropriate difficulty according to examinees' responses, thereby enabling individualized testing. In the current era of artificial intelligence and data informatization, CAT is a major direction in the development of testing technology research. In the history of the development of computerized adaptive testing, many researchers have previously provided certain summarized overviews of the measurement ideas and

development course of CAT, including Green (1970), Weiss and Bezt (1973), Weiss (1974, 1985), Killcross (1976), Weiss and Vale (1987), and Meijer and Nering (1999). These works mainly introduced the development of adaptive testing (computerized adaptive testing) during a particular period from the 1950s to the mid-1980s

1–6

. Chang (2015), Zhang Huahua and Cheng Ying (2005), Zhang Qinghua et al. (2007), and other researchers summarized CAT measurement technology from the mid-1970s to 2015

7–9

. However, these works have the following shortcomings in their discussion of the development process of the theory and technology of adaptive testing and computerized adaptive testing: (1) these works only discuss a single stage in the development of adaptive testing or CAT; (2) they discuss only part of the topics or measurement techniques of adaptive testing or CAT, without describing the developmental threads of adaptive testing and CAT testing; (3) in these works, only some researchers have examined the development of adaptive testing···

---

* Jian Xiaozhu, School of Public Policy and Administration, Nanchang University; Professor, master' s supervisor; research direction: educational and psychological measurement; email: jianxiaozhu2003@126.com; Funding projects: Regional Program of the National Natural Science Foundation of China (No. 32360204). Project title: Theoretical Hypotheses, Measurement Techniques, and Applied Testing of the Logistic Weighted Model for Psychological and Educational Testing. Later-stage funded project of the National Social Science Fund of China (No. 21FJKB021): Theory and Technology of Computerized Adaptive Testing.

a rough division; for example, Killcross (1976) divided adaptive testing into two stages, with around the end of the 1960s as the temporal boundary. [5] Weiss (1985) divided adaptive testing into three forms [4]: the first adaptive test (the Binet–Simon test), other forms of adaptive testing, and IRT-based adaptive testing. [4] However, Weiss (1985) did not discuss time boundaries for these three forms of adaptive testing, nor did he elaborate on the measurement characteristics of each stage. In addition, Weiss and Yan (2024), and Weiss and Sahin (2024, pp. 18–39) discussed multiple CAT case studies in the historical development of CAT and their characteristics [10, 11], but did not discuss, according to a logical thread, the internal logic and landmark contributions in the development of CAT measurement theory and measurement technology; nor did they divide the developmental stages by time, and thus they were unable to summarize the overall process and appearance of CAT' s historical development.

In the past, when the above researchers discussed the development of CAT testing, they generally believed that at the end of the 1960s (around 1970), adaptive

testing theory and measurement technology began to achieve breakthrough development. This article likewise takes the end of the 1960s as an important temporal boundary, dividing the development of adaptive testing before 1968 into two stages: the embryonic stage of adaptive testing (before 1942) and the early stage of adaptive testing (1942 to 1968); and dividing the development of adaptive testing in and after 1968 into three stages: the stage of theoretical breakthroughs in adaptive testing (1968 to 1975), the initial development stage of computerized adaptive testing (1975 to 2007), and the current development of computerized adaptive testing (after 2007).

This article will discuss the historical development process of adaptive testing and computerized adaptive testing from several major aspects: the theoretical development of adaptive testing and computerized adaptive testing, testing technology, and test formats.

2 Embryonic Stage of Adaptive Testing (1905 to 1942)

The French physicians Binet and Simon published the first edition and the revised edition of the Binet-Simon intelligence scale in 1905 and 1908, respectively. The Binet-Simon intelligence test arranged all items according to difficulty, from easy to difficult; examinees answered in order of difficulty; and test termination was determined according to the examinee' s responses. The Binet-Simon intelligence test proposed the concept of mental age as a measurement scale, used to express children' s level of intelligence and the difficulty level of items, thereby ingeniously placing examinee intelligence scores and test-item difficulty on the same scale (i.e., the mental-age scale). Therefore, the Binet-Simon intelligence test has been recognized by researchers as the earliest form of adaptive testing, for example by Weiss et al. (1973) [2] and Moses (2017) [12]; this is also the view recognized by the website of the International Association for Computerized Adaptive Testing (IACAT, 2025) [13]. The examinee ability scores and test-item difficulty of the Binet-Simon intelligence test were formally established on the mental-age scale, which is a key characteristic of measurement in adaptive testing. However, the method for establishing the mental-age scale was relatively crude and imprecise, relying mainly on methods of classical measurement theory to calculate item difficulty and examinees' mental age. Item difficulty was determined according to the pass rate: if 80% of normal children in a certain age group could pass an item, then that item was assigned to that age group. The highest age-group items that an examinee could pass were regarded as his or her mental age; for example, a child who passed all items for the 8-year-old group, regardless of actual chronological age, had a mental age of 8 years.

The Binet-Simon intelligence test administered testing in an adaptive-testing manner. At that time, research on the Binet-Simon intelligence test became popular worldwide. Many researchers carried out applied measurement concerning the Binet-Simon intelligence test, and some researchers further examined the principles of the testing method.

discussion. In Thurstone' s (1925) analysis of test data from the Binet-Simon intelligence test, he plotted the curve between the percentage of correct responses to items and the age scale[1]; this curve is precisely the item characteristic curve, revealing the quantitative relationship between item difficulty and the examinee' s ability level, and containing the basic theoretical assumptions of the "item characteristic curve" in item response theory.

Symonds (1929), under the ideal normal distribution, carried out mathematical derivations of test formulas and obtained several inferences[2], including: (1) the items that can most accurately measure a person' s ability are those on which the examinee can achieve a fifty-percent likelihood of answering correctly. (2) The test that can most accurately measure a person' s ability is one in which all items have difficulties such that the examinee can complete them with a 50% accuracy rate. These two propositions are the item-selection principles and testing goals for computerized adaptive testing: first, when selecting items, one should choose those for which the examinee has a possible probability of 50% of answering correctly; second, the testing goal of computerized adaptive testing is that, at the end of the test, the examinee' s accuracy rate on all items in the test should be around 50%.

In sum, in the embryonic stage of adaptive testing, the form of adaptive testing at this stage was mainly the Binet intelligence test. The development of measurement theory for adaptive testing in this stage is mainly reflected in: (1) adaptive testing (the Binet-Simon test) used mental age to place item difficulty and examinee ability on the same scale, but the computation of mental age relied on the calculation methods of classical measurement theory, which were crude and imprecise; (2) two inferences were proposed as the basis for selecting items in adaptive testing: the items that can most accurately measure a person' s ability are those on which the examinee can achieve a fifty-percent likelihood of answering correctly; the test that can most accurately measure a person' s ability is one in which all items have difficulties such that the examinee can complete them with a 50% accuracy rate. (3) A "item characteristic curve" graph was plotted for the relationship between the percentage of correct responses to items and the age scale, which contained the basic theoretical assumptions of item response theory.

3 Early Stage of Adaptive Testing (1942 to 1968)

3.1 Basis for Dividing the Early Stage of Adaptive Testing

This paper takes 1942 as the temporal dividing point between the early stage of CAT development and the preceding stage, mainly on the basis of the following two references:

(1) Drawing on previous researchers' views on the timeline of adaptive-testing

[1]14
[2]15

development. Killcross (1976) argued that adaptive testing emerged and developed after World War II, and counted a total of 17 relevant papers on adaptive testing from 1944 to 1967[3]. Weiss (1985) believed that it was not until around 1950 that researchers began to expand the scope of adaptive-testing applications from intelligence tests to school academic-ability tests, including vocabulary tests and mathematics-ability tests[4]. The Educational Testing Service (ETS) in the United States also began, after World War II, to attempt to use adaptive testing in the field of educational testing; ETS research reports on forms of adaptive testing stimulated many other researchers to further explore adaptive-testing methods.

(2) In 1942 and 1943, new forms of adaptive testing and early forms of IRT models were proposed. For example, Spache (1942), when conducting the Stanford-Binet test, adopted a continuous serial testing method and an age-level-by-age-level testing method[5], selecting items according to whether the examinee answered correctly or incorrectly; this was an improvement on the Binet-Simon form of adaptive testing. Hutt (1947) applied the Stanford-Binet Intelligence Test (Revised Edition) and compared two testing methods (a continuous test in order of difficulty, and an adaptive test that selected the next more difficult or easier item according to the examinee' s responses)[6]. In addition, Ferguson (1942) and Lawley (1943) proposed, in measurement theory, probabilistic models of the functional relationship between item difficulty and examinee ability parameters; these were embryonic forms of early IRT models.[7]

Next, this paper discusses research on the early stage of adaptive testing in two parts: theoretical research and empirical research (including simulation research).

3.2 Theoretical Development in the Early Stage of Adaptive Testing

The development of measurement theory in the early stage of adaptive testing mainly involved two aspects:

The first aspect was that researchers proposed early prototypes of IRT models in which item difficulty parameters and examinee ability parameters were placed on the same scale. Under the premise of an ideal normal distribution, Ferguson (1942) derived estimation formulas for item discrimination parameters and difficulty parameters, deduced the relationship between item parameters under classical test theory and item response theory, and established a probabilistic model of the functional relationship between item difficulty and examinee ability

[3] 5
[4] 6
[5] 16
[6] 17
[7] 18, 19

parameters, placing item difficulty parameters and examinee ability parameters on the same scale. This was an important milestone in the development of measurement theory for adaptive testing.[18] Building on Ferguson' s (1942) research, Lawley (1943) further proposed a probability formula for an examinee' s correct response to an item[19]. This formula used the discrimination and difficulty parameters of an item to express the cumulative-function probability that an examinee at a particular ability level would pass that item; this cumulative function placed item difficulty parameters and examinee ability parameters on the same scale, and may be regarded as an early prototype of the IRT model. Under the ideal normal distribution curve, Lord (1952) used the normal cumulative ogive curve model to place examinee ability parameters and item difficulty parameters within the same measurement model—that is, on the same scale—which was an important breakthrough in IRT measurement theory[20].

The second aspect was that researchers proposed the theoretical basis for selecting items of corresponding difficulty according to the examinee' s ability level during the adaptive testing process. Hick (1951), deriving and discussing from the perspective of information theory, argued that, considering testing efficiency and information, an individualized test of intelligence should begin with an item for which the examinee has an overall probability of 0.5 of answering correctly. If the examinee answers correctly, the next item should be selected such that the examinee has a 0.5 probability of answering it correctly[21]. Hick' s idea that an examinee should have a 0.5 probability of answering an item correctly became the principal principle underlying later adaptive test item selection. The Fisher maximum-information and *b*-matching strategies subsequently proposed by researchers for CAT item selection are fundamentally consistent with the principle of a 0.5 probability of correct response.

3.3 Empirical Applications and Simulation Research in the Early Stage of Adaptive Testing

Weiss (1985) noted that it was not until 1950 that researchers began to extend the application scope of adaptive testing from intelligence tests to school academic ability tests, including vocabulary tests and mathematical ability tests[6]. In the 1950s, many researchers conducted empirical or simulation studies of adaptive testing and proposed various forms of adaptive tests. These adaptive tests included: the sequential analysis test (Cowden, 1946)[22], the two-stage adaptive test form (two-stage test, Angoff, & Huddleston, 1958)[23], the pyramidal adaptive test (pyramidal test, Krathwohl, & Huyser, 1956)[24], and the branched testing form (branched testing, Bayroff, 1964)[25]. Weiss and Betz (1973) classified these adaptive testing forms according to testing stages[2], grouping them uniformly into two broad categories: two-stage adaptive testing strategies (two-stage strategies) and multi-stage adaptive testing strategies (multi-stage strategies).

Angoff and Huddleston (1958) proposed that the two-stage adaptive testing strategy is essentially a kind of multistage design and belongs to computerized

multistage adaptive testing (MST)[23]. In addition, Cowden (1946) applied the sequential sampling procedure to conduct an empirical study of a class final examination, classifying examinee achievement levels by means of adaptive testing; in terms of test form, it was a two-stage adaptive test[22]. Moonan (1950) applied the sequential analysis test to scholastic

In calibration ability tests [26], adaptive testing was used to classify examinees. These adaptive tests using sequential analysis for measurement were the earliest testing forms that used adaptive testing to classify examinees, and may be regarded as an early form of computerized classification testing.

3.4 Commentary on the Development of Testing in the Early Stage of Adaptivity

The scope of application of adaptive testing expanded from intelligence tests to school academic-ability tests, and a variety of adaptive testing formats were proposed to improve measurement techniques. However, academic-ability testing differs greatly from intelligence testing. Academic-ability tests contain a large number of items, and differences among item difficulty levels are not like those in intelligence-test items, where item difficulty can be defined simply by mental age. Although many researchers designed adaptive tests and continuously improved adaptive-testing formats, testing routes, and scoring rules, in academic-ability testing researchers generally used the pass rate P to define and describe item difficulty parameters, while estimates of examinee ability scores were comprehensively evaluated and determined according to such factors as the examinee' s testing route, the difficulty of the last item answered correctly, and the number of items answered correctly. Thus, although the design of such testing processes also belonged to adaptive-testing formats, examinee ability scores and item-difficulty parameters were not placed on the same scale. Therefore, when adaptive testing was applied to academic-level testing at that time, researchers proposed many adaptive-testing formats; what was urgently needed then was a measurement theory and measurement model to unify the research directions in adaptive testing at the time. Weiss et al. (1973) believed that, for adaptive-testing formats in the early stage, the appropriateness of item analysis and item selection methods was also controversial, and that classical test theory item methods might be inapplicable in adaptive-testing contexts [1]. Early researchers also recognized the difficulties of comparing the consistency of multiple testing formats in adaptive testing, as well as the difficulties regarding the consistency of examinee ability-level scoring rules and interpretations. Reflection on these issues and the proposed solutions promoted the subsequent development and application of item response theory measurement ideas and measurement methods. Lord (1968) proposed the viewpoint concerning the relationship between adaptive testing and measurement theory [27], which was a theoretical research exploration that elevated the development of the various adaptive-testing formats at that time.

4 Breakthrough Stage of Adaptive Testing (1968 to 1975)

4.1 Basis for Dividing the Breakthrough Stage of Adaptive Testing

This paper takes 1968 as the time boundary with the preceding stage and classifies the period from 1968 to 1975 as the breakthrough stage of adaptive testing, mainly on the following two bases: (1) many researchers generally believe that the late 1960s was an important time demarcation point in the development of adaptive-testing research. Killcross (1976) held that in the late 1960s, because computers began to be used in laboratories, this greatly promoted research and application of adaptive testing. [5] (2) During the period from 1968 to 1975, researchers proposed that adaptive testing needed guidance from measurement theory, and CAT research based on IRT models gradually began to become mainstream. Lord (1968) first proposed the view that testing and item selection in adaptive testing needed to be based on measurement theory. [27] Lord (1972) introduced the term item characteristic curve theory, and discussed and analyzed the relationship between adaptive testing and item characteristic curve theory (i.e., item response theory). In the study, an IRT model was used to estimate examinee ability and to select items of appropriate difficulty, placing examinee ability and item parameters on the same scale [28]. Researchers such as Zhang Huahua believed that placing examinee ability and item parameters on the same scale is a key feature of CAT testing [29]. Lord (1975) further explicitly proposed the viewpoint of calibrating item difficulty parameters on the examinee ability scale. [30] Other researchers also gradually began to design adaptive testing on the basis of item response theory as the measurement-theory foundation.

for example, Owen (1969, 1975) applied the IRT model in adaptive testing, using Bayesian estimation methods to estimate examinees' ability levels according to the magnitude of item-difficulty parameters and examinees' response patterns. [31, 32]

4.2 Progress in Theoretical Research During the Breakthrough Stage of Adaptive Testing

During the breakthrough stage of adaptive testing, Lord and Novick (1968) [33], Lord (1972) [28], and Wright and Panchapakesan (1969) [34] proposed and discussed the basic postulates, basic theoretical assumptions, and inferences of item response theory, as follows:

(1) Basic postulates. Lord and Novick (1968) discussed the basic assumptions (postulates) of latent trait theory (item response theory): in a testing situation, an examinee' s test performance can be interpreted by defining the examinee' s characteristics (traits) and estimating the examinee' s scores on these traits; moreover, the scores can be used to predict the

examinee' s response performance on test items. [34]

(2) Basic theoretical assumptions. Lord (1972) first discussed three prerequisite assumptions of item response theory (item characteristic curve theory) [28], namely the item characteristic curve function assumption, the local independence assumption, and the unidimensionality-of-ability assumption. Specifically: (a) the item characteristic curve function assumption requires testing whether the test data fit the IRT item characteristic curve function (IRT model); (b) the local independence assumption requires that the probability of an examinee answering one item correctly not be affected by his or her performance on other items. This requires that examinee responses not exhibit abnormal scoring phenomena, including examinee cheating, answer cueing among items, and similar phenomena. (c) The unidimensionality assumption of ability as a latent trait requires testing the unidimensionality assumption for test data. Once these three conditional assumptions hold, the IRT model can be applied to this set of test data. It should be noted in particular that the item characteristic curve function assumption is the basic theoretical assumption of IRT, whereas the local independence assumption and the unidimensionality-of-ability assumption are prerequisite or conditional assumptions for the IRT model to hold.

(3) Two basic inferences. Wright and Panchapakesan (1969) discussed two basic inferences of the IRT model [34]: (a) estimation of item parameters does not depend on (is independent of) the selection of the examinee-group sample; that is, when an item is administered to different examinee groups, the estimated item parameters for that item are the same. (b) Estimation of examinee ability parameters does not depend on (is independent of) the selection of test items; that is, when an examinee responds to different sets of items, the estimated ability parameter for that examinee is the same.

Because the normal distribution density integral function proposed by Lord (1952) required a large amount of integral computation in practical applications and was difficult to compute in practice, Lord and Novick (1968), through theoretical derivation and computation, found [33] that the absolute difference between the normal distribution density integral function and the Logistic function was less than 0.01. This provided a solid theoretical basis for subsequent researchers' choice of the Logistic function. Because the Logistic function is simplified and easy to compute, later researchers in CAT mainly used two- and three-parameter Logistic models.

4.3 Development of Test Forms During the Breakthrough Stage of Adaptive Testing

Weiss (1974) reviewed and summarized the various adaptive testing forms that appeared during the breakthrough stage of adaptive testing (1968–1974) [35], including tailored testing in the style of measuring an ability vector (Lord, 1971)

[36], programmed testing (Cleary, Linn & Rock, 1968a) [37], multilevel elastic testing (Lord, 1971) [38], measurement based on examinee test responses (Wood, 1970) [39], the stratified adaptive testing proposed by Weiss (1973) [40], and multistage adaptive testing forms. These testing forms differed considerably from one another in item-selection rules, ability-estimation methods, and test-termination strategies. Among them, the adaptive testing form of multistage testing ···

The forms are divided into fixed-branch adaptive testing (Wood, 1969) [41] and variable-branch adaptive testing (also called mathematical strategy testing).

Variable-branch adaptive testing can be further divided into the Bayesian adaptive testing form (Owen, 1969; Novick, 1969) [30, 42] and the maximum-likelihood model adaptive testing form (Urry, 1970) [43]. Subsequent researchers have generally recognized two relatively general models: the Bayesian adaptive testing form and the maximum-likelihood adaptive testing form. Moreover, since 1975, CAT researchers have basically carried out improvements and extensions under these two adaptive testing models.

(1) Bayesian-strategy adaptive testing. The Bayesian strategy adopted in adaptive testing is based on the application of Bayes' theorem in adaptive testing (Owen, 1969; Novick, 1969). The general steps of applying the Bayesian method to adaptive testing are as follows. First, at each stage of the test, according to the information available about the examinee, the examinee' s ability and standard error are estimated in advance before testing. Second, an item is selected from the item bank; this item is the one that will reduce, to the greatest extent, the uncertainty of the examinee' s ability estimate. If a certain examinee is selected to take the test, the selected item is usually the item whose difficulty is closest to the examinee' s estimated ability level. After the selected item has been administered, Bayes' theorem is used to combine the prior ability estimate with the examinee' s response information (correct or incorrect), yielding a posterior ability estimate, while the error of the new posterior ability estimate is also calculated. If the error reaches a sufficiently small value, the test can be terminated; if the estimation error of the posterior ability estimate is smaller than the preset testing error, testing continues until it becomes smaller than the preset testing error.

(2) Maximum-likelihood-model adaptive testing. Urry (1970) proposed an adaptive testing strategy based on the maximum-likelihood method. The operating mode of the maximum-likelihood adaptive testing process is similar to the Bayesian process, but the mathematical principles are different. After the examinee answers one item correctly and another item incorrectly, the maximum-likelihood equation can be solved, and the ability estimate and its standard error can be obtained. The next selected item is the item in the item bank whose difficulty level is closest to the examinee' s estimated ability value. According to the examinee' s response to each item, the maximum-likelihood procedure generates an ability es-

timate and its standard error. When the standard error of the ability estimate reaches a predetermined value, the test can be terminated.

5 Initial Development Stage of CAT (1975 to 2007)

5.1 Basis for Dividing the Initial Development Stage of CAT

Taking 1975 as the time boundary between the initial development stage of CAT and the preceding stage, the division is based on the following three landmark changes:

(1) The first landmark: theoretical research and practical application of computerized adaptive testing were mainly based on item response theory. After 1975, CAT research and application gradually abandoned the classical measurement-theory approach of using difficulty as the basis for ordering item difficulty and selecting testing paths, and gradually took item response theory as the measurement-theoretical basis. Lord (1977) formally renamed item characteristic curve theory as item response theory [44], and item response theory was proposed as a formal testing theory. The computerized adaptive testing discussed in this paper is all based on item response theory, that is, IRT-based CAT.

(2) The second landmark: computers began to be applied to adaptive testing research, replacing the previous manually interactive adaptive tests that relied on paper-and-pencil testing, which promoted CAT theoretical research and practical application. Weiss and Bock (1983) stated in the preface [45] that in the mid-1970s, minicomputers provided interactive computing capability in research laboratories and real testing environments; and in the few years after 1975, research covered and surpassed the test development and psychometrics of adaptive testing over the previous more than 60 years.

related fields. After 1975, adaptive testing gradually shifted from the traditional paper-and-pencil response format to computer-based testing, and used IRT as its measurement-theory foundation.

(3) Milestone Three: the convening of CAT symposia. In 1975, researchers at the University of Minnesota held the first symposium on computerized adaptive testing research (Clark, 1976) [48]. The conference was named "Computerized Adaptive Testing," and this name also began to be widely used in the research literature from that conference onward. From 1975 to 1982, the University of Minnesota organized four CAT symposia. These four CAT symposia greatly promoted research on the technology and applications of computerized adaptive testing.

From 1975 to 2007, item response theory made substantial progress in measurement theory and measurement technology, including methods for estimating item parameters, estimating examinee ability, test and item information, test

equating, and related measurement theories and techniques. The following discussion addresses two aspects: the development of measurement theory and the development of measurement technology.

5.2 Development of Measurement Theory in the Initial Stage of CAT

In terms of the development of CAT measurement theory, Lord first used the term "item response theory" in an informal report in 1976 [47]. Lord (1977) formally renamed item characteristic curve theory as item response theory [44]. Warm (1978) summarized the measurement ideas concerning IRT among previous researchers, summarized the three basic assumptions of item response theory, and so on; he also changed the term "item characteristic curve" in Lord's (1977) work to "item response curve," the term "item characteristic function" to "item response function," and the term "item characteristic curve theory" to "item response theory," taking it as the measurement-theory basis for adaptive testing and computerized adaptive testing [48]. In 1980, Lord published the first monograph on item response theory [49], which researchers generally regard as marking the basic maturation of item response theory. Later researchers gradually adopted the term "item response theory" and used it as the measurement-theory basis for computerized adaptive testing. Hambleton et al. (1991) summarized previous research ideas and provided a relatively comprehensive overview and formation of the basic theoretical system of item response theory, including the basic axioms of item response theory, the basic theoretical assumptions of IRT, and two basic inferences [50, pp. 7–9].

5.3 Development of Measurement Technology in the Initial Stage of CAT

At this stage, IRT-based CAT measurement formats were the mainstream, while other measurement formats not based on IRT were gradually eliminated by researchers. After 1975, the three main components of adaptive testing (namely item selection strategy, ability estimation, and termination strategy) began to apply item response theory in theoretical or empirical research, and two main models gradually took shape: the Bayesian model and the maximum-likelihood model. van der Linden and Glas (2000) discussed these two models [51]. (1) Bayesian model. Owen (1969, 1975) proposed using the least-variance method based on the Bayesian posterior distribution as the item selection strategy, using Owen's posterior ability estimation method to estimate examinee ability, and using the criterion that the minimum posterior variance is smaller than a specified value as the test termination strategy. (2) Maximum-likelihood model. Urry (1970) was the earliest to propose the maximum-likelihood model; Lord (1977) further improved the maximum-likelihood model [48], proposing that the Fisher information function be used as the item selection strategy in CAT, that the maximum-likelihood estimation method be used to estimate examinee ability, and that the test information reaching a specified value be used as the

test termination strategy. Subsequent CAT research has basically improved, combined, or extended these several main components of the two CAT models.

The measurement technologies for each link in the CAT testing process also gradually developed, including CAT item selection strategies, item exposure-rate control, ability estimation methods, termination strategies, and online calibration of CAT item parameters. These measurement technologies are discussed only briefly below.

(1) Development of measurement techniques for CAT item-selection strategies. Lord (1977) proposed using the Fisher information strategy in CAT tests to select items[8], that is, selecting items with the maximum item information according to the examinee' s ability for testing. This was the first item-selection strategy algorithm based on an IRT model. Later, according to the needs of actual CAT testing situations, researchers further proposed derivative strategies of the Fisher information strategy, as well as the K-L information-function item-selection strategy, the α-stratified item-selection strategy, the b-matching item-selection strategy, the Bayesian item-selection strategy, and derivatives or comprehensive strategies of these strategies. To realize multiple test-objective constraints in CAT, researchers proposed mathematical models that simultaneously control multiple constraints, such as the weighted-deviation model, the optimal-index model, the shadow-test method, and the Monte Carlo simulation method. Jian Xiaozhu et al. (2014) provided an overview of these CAT item-selection strategy methods[9].

(2) Development of measurement techniques for item-exposure-rate control. For item-exposure-rate control, the simplest method is to adopt randomization methods, including the M-M random strategy, the 1-0 logical random method, the random-program method, and variants of randomization methods such as the progressive method. Sympson et al. (1985) proposed a conditional-probability method for CAT item control (abbreviated as the SH method)[10]. Subsequent researchers further improved the SH method for item-exposure-rate control, or incorporated statistical methods, gradually forming four major types of item-exposure-control methods: the SH conditional-probability method and its variants; the item-eligibility method and its variants; the multiple maximum-exposure-rate method; and the α-stratification method and its variants. Li Mingyong et al. (2010) reviewed these item-exposure-rate control methods[11].

(3) Development of measurement techniques for examinee ability estimation methods. CAT ability estimation methods are mainly divided into three types and their variants: (a) maximum-likelihood estimation methods and their variants. Lord (1953) was the first to use maximum-likelihood estima-

[8] 48
[9] 52
[10] 53
[11] 54

tion to estimate examinee ability[12]. (b) maximum a posteriori estimation methods and their variants; (c) Bayesian expected a posteriori estimation methods and their variants. Owen (1975) extended the Bayesian posterior estimation method to the estimation of examinee ability in adaptive testing[13]. In recent years, some researchers have proposed comprehensively applying different ability estimation methods at different stages of CAT or under different testing situations.

(4) Development of measurement techniques for test-termination strategies. Among termination strategies of variable test-length types, there are mainly three categories of measurement-index reference bases: (a) termination strategies based on the measurement standard error, including the ability confidence-interval method based on the measurement standard error; (b) termination strategies based on Bayesian estimation methods, mainly including the minimum expected posterior variance criterion proposed by Owen (1969)[14] and the posterior standard deviation criterion proposed by Bock and Mislevy (1982)[15]; (c) termination strategies based on the likelihood-ratio method. Some variants have been derived from the above three categories of termination strategies. In addition, researchers have also proposed other termination strategy methods, including the minimum-information criterion for what the item bank can provide, the convergence criterion for the examinee ability estimate θ, stratified termination strategies, and CAT termination strategies such as Bayesian decision-theory rules.

(5) Development of measurement techniques for online calibration of CAT item parameters. Stocking (1988) first used conditional maximum-likelihood estimation methods to conduct online calibration of item parameters, including Method A and Method B[16]. Wainer and Mislevy (1990) further proposed the marginal maximum-likelihood estimation EM algorithm; the parameter estimation of this method is comparatively superior[17].

It can be seen from the above that, from 1975 to 2007, CAT gradually established item response theory as the basic testing theory, and gradually constructed the theoretical system of IRT. The measurement techniques for all links of the CAT testing process were also gradually improved and developed. During this period

During this period, researchers compiled and wrote a number of monographs on CAT, including Weiss and Bock (1983) [45]; Wainer et al. (1990) [59]; Sands et al. (1997) [60]; Drasgow and Olson-Buchanan (1999) [61]; and van der Linden and Glas (2000) [62].

[12]55
[13]31
[14]30
[15]56
[16]57
[17]58

5.4 Practical application cases in the initial development stage of CAT

In the initial development stage of CAT, CAT research gradually shifted from laboratory research to large-scale practical application. Research and design of the U.S. Armed Services Vocational Aptitude Battery began in 1979, and in 1985 the U.S. military' s computerized adaptive testing version of the Armed Services Vocational Aptitude Battery (ASVAB-CAT) officially began large-scale testing and application [60]. Large-scale operational CAT cases also included the following: in 1993, the U.S. Graduate Record Examination (GRE) began adopting a computerized adaptive testing format; in 1994, the U.S. nursing licensure examination adopted computerized adaptive testing; in the autumn of 1997, the U.S. Graduate Management Admission Test (GMAT) began adopting computerized adaptive testing. In addition, from 1991 to 2007, many U.S. professional qualification examinations—including the Architect Registration Examination, medical licensure examinations, U.S. teacher certification examinations, judicial examinations, practicing physician licensure examinations, and Microsoft software programmer certification examinations—gradually began to adopt CAT testing formats.

Quite a few research institutions or universities in China carried out applied research on CAT operational testing. For example, (1) Tu Shuqing et al. (2003) reported several operational CAT systems [63], including a CAT testing system for an adaptive mathematics intelligence-level examination developed in 1987, and an adaptive test for behavioral-scenario judgment of Party and government leading cadres developed in 2003. (2) The Basic Competence Test for Junior High School Students in Taiwan, China, began adopting a computerized adaptive testing format in 2002.

6 Current development stage of CAT (after 2007)

6.1 Basis for dividing the current development stage of CAT

The current development stage of CAT takes 2007 as the temporal boundary. This paper mainly refers to two indicators:

(1) Indicator 1: Around 2007, in order to meet practical needs, CAT research developed from the general CAT testing format into three CAT measurement-technology directions. Based on item response theory and the key features of CAT (i.e., that item difficulty parameters and examinee ability parameters are established on the same scale), Jian Xiaozhu and Zhang Minqiang (2024) summarized the many diverse CAT testing formats since 2007 into three CAT measurement-technology directions [64], namely cognitive-diagnostic computerized adaptive testing, computerized multistage adaptive testing, and computerized classification testing. The main time point at which research and applied practice in these three

measurement-technology directions developed was approximately around 2007.

(2) Indicator 2: In 2007, the CAT Research Center at the University of Minnesota organized a CAT symposium. This was the first such meeting after the 1982 CAT symposium had ceased, reconvened after an interval of 25 years. Beginning in 2007, the International Association for Computerized Adaptive Testing (IACAT) has held a CAT symposium once every one or two years, and these symposia have greatly promoted the development of CAT research. In 2010, the International Association for Computerized Adaptive Testing (IACAT) was established at the University of Minnesota in the United States, and it founded a professional journal for CAT research—*Journal of Computerized Adaptive Testing*.

6.2 Three directions in the development of CAT measurement technology

Since 2007, in accordance with the needs of practical CAT testing applications, the development of CAT measurement technology has shown three CAT measurement-technology directions: IRT-based cognitive-diagnostic computerized adaptive testing (cognitive-diagnostic CAT), computerized multistage adaptive ···

testing (MST), and computerized classification testing (CCT).

Jian Xiaozhu and Zhang Minqiang (2024) classified CAT studies that have appeared since 2007 and that involve a complex variety of test names and test forms, summarizing them into three main directions in the development of CAT testing technology: cognitive diagnostic computerized adaptive testing, computerized multistage adaptive testing, and computerized classification testing. (1) Cognitive diagnostic computerized adaptive testing (CD-CAT). The characteristic of this testing form is that the measurement model includes a cognitive diagnostic model and an IRT model, or a fusion model of a cognitive diagnostic model and an IRT model. Its measurement goals are both to estimate examinee ability and to conduct cognitive diagnosis, thereby providing a reference for teaching practice and learning remediation. Its Chinese names include cognitive diagnostic computerized adaptive testing, computerized adaptive diagnostic testing, and cognitive diagnostic computerized adaptive testing; its English name is cognitive diagnostic computerized adaptive testing. (2) Computerized multistage adaptive testing (MST). The characteristic of this testing form is that, in a CAT test, items are selected according to testlet modules (test modules), and the test is administered in multiple stages. Its measurement goals take into account the advantages of both general CAT and traditional testing, enabling it to effectively balance measurement accuracy and measurement precision, as well as other test-constraint objectives. Its Chinese names include computerized multistage adaptive testing, computerized multistage adaptive testing, multistage adaptive testing, computerized adaptive sequential testing, and cognitive di-

agnostic multistage testing, among others; its English names include computer adaptive multistage testing, computer-adaptive sequential testing, bundled multistage adaptive testing, multistage testing, multistage adaptive testing, and adaptive multistage testing. (3) Computerized classification testing (CCT). The characteristic of this testing form is that, in the main testing procedures—such as CAT item-selection strategies and termination strategies—it proceeds according to cut scores for examinee ability. Its measurement goal is to achieve relatively high measurement precision at the cut scores for examinee ability and to classify examinee ability; it is especially suitable for vocational qualification examinations and pass/fail assessment tests. Its Chinese names include computerized classification testing, computerized adaptive classification testing, and computerized mastery testing; its English names include computerized classification testing, mastery adaptive test, computerized adaptive classification test, adaptive mastery testing, computerized mastery adaptive test, mastery testing with CAT, and Pass-Fail CAT. The characteristics and distinctions of these three directions are organized and listed below in tabular form; see Table 1.

Table 1 Measurement objectives and testing formats of three directions in the development of CAT testing technology

Test name	Measurement objective	Testing format	Characteristics of the testing technology	Typical example
IRT-based cognitive diagnostic computerized adaptive testing Foreign-language name: IRT-Based cognitive diagnostic computerized adaptive testing, IRT-Based CD-CAT	Estimate the examinee' s ability and diagnose the examinee' s knowledge mastery status	The testing process still selects items in the form of one item at a time, consistent with general CAT	Although based on both IRT models and cognitive diagnostic models, the specific algorithms for item-selection strategies, termination strategies, ability estimation, etc., differ from those of general CAT	The "Yincaiwang" platform developed in Taiwan, China, based on CD-CAT technology

Test name	Measurement objective	Testing format	Characteristics of the testing technology	Typical example
Computerized multistage adaptive testing Foreign-language name: computer multistage adaptive testing, MST;	It combines the advantages of traditional testing and computerized adaptive testing	In the testing process, items are selected and administered in the form of item-group modules	The assembly strategies and routing rules for test modules differ from those of general CAT	The GRE-CAT test revised in 2011; Chinese Proficiency Grading Test
Computerized classification testing Foreign-language name: Computerized Classification Testing, CCT	Classify examinees according to ability cut scores	In the testing process, items are selected with ability cut scores as references	Item-selection strategies and termination strategies are specifically based on ability cut scores	U.S. nursing licensure examination

IRT-based cognitive diagnostic computerized adaptive testing, computerized multistage adaptive testing, and computerized classification testing—these three forms of CAT testing technology—differ to some extent from the measurement-technology form of general CAT, but in essence they still belong to IRT-based CAT measurement formats and conform to the key characteristics of CAT, namely, that item difficulty parameters and examinee ability parameters are established on the same scale. Jian Xiaozhu and Zhang Minqiang (2024) summarized the measurement theories, measurement technologies, and practical application situations of these three measurement directions in their article[^64]. Regarding the three directions of CAT measurement technology development, many foreign researchers hold similar views. In their monograph, Yan, von Davier, and Weiss (2024)[^65]: (1) divide CAT formats into general CAT testing formats and computerized multistage adaptive testing, and note that computerized multistage adaptive testing can be combined with cognitive diagnosis; (2) regard computerized classification testing as an independent CAT testing format, while computerized multistage adaptive testing can also incorporate classification algorithms from computerized classification testing to classify examinees. In their monograph, Weiss and Sahin (2024) distinguish CAT into general CAT

testing formats, computerized classification testing, and other types of CAT (including computerized multistage adaptive testing, cognitive diagnostic CAT, etc.)[^66].

6.3 Development and application of CAT measurement technology in China at the current stage

Before 2006, there were relatively few CAT testing studies by domestic researchers, most of which were review studies and applied case studies. Since 2006, domestic research on improvements and innovations in CAT testing technology has gradually increased; the period around 2006 represents a relatively clear temporal dividing point. After 2006, domestic researchers proposed multiple improvement methods concerning CAT item-selection strategies, item exposure-rate control, ability estimation methods, test termination strategies, online calibration of item parameters, and other aspects. Only some representative studies are listed here. For example, Chen Ping, Ding Shuliang, Lin Haijing, and Zhou Jian (2006), and Dai Haiqi, Chen Dezhi, Ding Shuliang, and Deng Taihua (2006) proposed multiple item-selection strategies for the GRM model[^67,^68]; Cheng Xiaoyang et al. (2011) proposed an item-selection strategy introducing an exposure factor[^69]; Zhu Longyin et al. (2008) proposed an exploration combining test termination strategies with stratified item-selection strategies[^70]; Chen Ping and Ding Shuliang (2008) proposed a study exploring a CAT testing format that allowed returning to revise answers[^71]; Chen Ping (2016) proposed an improved method for online calibration technology for item parameters in new items[^72]. Among the three directions of CAT measurement technology development, domestic researchers have conducted more research in the direction of cognitive diagnostic CAT measurement technology; Liu Yan et al. (2017) on

A bibliometric analysis was conducted on research into cognitive diagnostic CATs[^73]. Yang Tao, Huang Shengshun, and Xin Tao (2015), and Wang Yuxiang, Luo Zhaosheng, and Wang Rui (2015) reviewed the concepts, test-design ideas, and measurement techniques of previous studies on computerized multistage adaptive testing[^74,^75]. Jian Xiaozhu and Chen Ping (2020) summarized and discussed the principles and testing techniques of computerized classification testing, focusing on item-selection strategies and test-termination strategies for computerized classification testing[^76]. Ren He and Chen Ping (2021) proposed two new stopping rules for multidimensional computerized classification testing[^77].

There are also multiple domestic application cases of CAT tests, including: (1) the Hanyu Shuiping Kaoshi for Overseas Chinese Students, sponsored by the National Office for Teaching Chinese as a Foreign Language, adopts the CAT test format (computerized multistage adaptive testing format)[^78]; (2) the Chinese Military Medical and Psychological Selection System compiled by the Fourth Military Medical University uses the CAT test format for selecting and screening new recruits[^79].

6.4 Current Frontier Research on the Application of AI Technology to CAT Testing

Over the past five years, the hottest frontier in CAT testing-technology research has undoubtedly been the application of artificial intelligence (AI) to CAT measurement technology. At present, the integration of CAT measurement technology with artificial intelligence (AI) is mainly reflected in two aspects: automatic item generation and online calibration for CAT item banks, and CAT item-selection strategies. (1) Applying AI technology to automatic item generation and item-parameter calibration for CAT item banks. Wang Lei (2023) discussed application cases of artificial-intelligence technology (including ChatGPT language models) in automatic item generation, as well as automatic item generation for different item types[^80]. Chen Zhe et al. (2024) introduced an item-authoring system that applies large-scale AI language models to self-study examinations in higher education, and described its applied process technologies such as automatic item generation, automatic review, and test-paper generation[^81]. Weiss and Sahin (2024) discussed the issue of applying AI to automatic item generation for CAT item banks, arguing that items generated with the help of AI can serve as drafts, but ultimately still need to undergo manual review and evaluation by subject-matter experts (SMEs)[^67, pp.325-345]. Han Yiting et al. (2025) reviewed how previous researchers used AI model technologies to realize automatic item generation and optimize the quality and data structure of items in CAT item banks[^82]. (2) Improving CAT item-selection strategies through AI methods. Chen et al. (2022) combined AI model algorithms with deep learning to optimize item selection in CAT, balancing the information amount and knowledge coverage of test items and avoiding excessive repeated exposure of a certain knowledge point and its corresponding items[^83]. Wang et al. (2024) improved CAT item-selection strategy technology under an AI model by combining reinforcement-learning methods, and proposed an item-selection strategy based on a deep Q-network method (DQN). They found that, under the new item-selection strategy method, test simulation indicators RMSE and MAE were both superior to several traditional item-selection strategies[^84].

In terms of practical applications combining AI and CAT testing, the Duolingo English Test (DET) is a relatively typical case. The Duolingo English Test develops CAT tests by combining machine learning and natural language processing on the basis of AI technology. Cardwell et al. (2024)[^85], Burstein et al. (2022)[^86], and Langenfeld et al. (2022)[^87] reported applications of artificial intelligence (AI) in the development and testing process of Duolingo English tests, including the use of AI in item development (using AI technology for automatic item generation), testing-process management (online calibration of item parameters and routing rules in adaptive testing), automatic scoring of examinees' oral responses, and the combination of AI proctoring with human proctoring.

At present, there are great differences among researchers in their use of artificial-

intelligence (AI) technology in CAT research and in their application of AI technical algorithms. The author has searched and reviewed the literature on the combination of CAT and artificial-intelligence technology, and the summarized findings are as follows: (a) the technical directions and algorithms by which different researchers apply AI technology to CAT vary greatly, including different technical routes such as machine learning, deep learning, and neural-network training models; and no generally recognized, typical test form or technical algorithm that combines AI technology with CAT testing technology has been found.

; (b) in these studies on the application of AI technology to CAT, there are no distinctive measurement objectives, nor are there hallmark characteristics in the form of measurement;

(c) there is no technical research or practical application case in which a particular AI technology can be applied throughout the entire CAT testing-technology process.

Therefore, the authors believe that the current application of AI technology to CAT testing technology is basically still in an early exploratory stage. This paper argues that if an integrated model combining AI and IRT can be fully applied to each testing process of CAT, or if a certain AI technical algorithm can be applied throughout the entire CAT testing process, then the direction of measurement technology for applying AI to CAT may be defined as the fourth developmental direction of IRT-based CAT, namely AI-CAT. If AI technology is applied only to a certain stage or partial stage of the CAT testing-technology process, and the measurement objectives and measurement forms of the application of AI technology show no distinctive characteristics, then the direction of measurement technology for applying AI to CAT is difficult to define as the fourth developmental direction of CAT measurement technology. Of course, whether it can become the fourth developmental direction of IRT-based CAT measurement technology in the future still awaits researchers' exploratory breakthroughs in technologies integrating AI and CAT.

7 Conclusion

Based on the developmental process of CAT test forms and the measurement theories on which they rely, this paper divides the development of CAT testing into five stages. The measurement forms and measurement characteristics of the five historical stages of CAT test development are summarized in tabular form here; see Table 2:

Table 2 Summary of the Five Historical Stages in the Development of Adaptive Testing and Computerized Adaptive Testing

Measurement theory	Testing format	Test characteristics	Characteristics of adaptivity	Scope of test application
Embryonic stage of adaptive testing (1905–1942)	Classical measurement theory	Paper-and-pencil test	Using mental age as the measurement scale and as the basis for item selection	Binet–Simon intelligence test
Early stage of adaptive testing (1942–1968)	Classical measurement theory	Paper-and-pencil test	Preset all testing paths Designed multiple adaptive testing forms with fixed testing paths	Expanded from intelligence tests to school academic ability tests
Breakthrough stage of adaptive testing (1968–1975)	Transition from classical measurement theory to item response theory	Paper-and-pencil tests began to shift toward computerized testing	Began to estimate examinee ability and item parameters based on IRT models, and used these as item-selection strategies	School academic ability tests

Measurement theory	Testing format	Test characteristics	Characteristics of adaptivity	Scope of test application
Initial development stage of CAT (1975–2007)	Item response theory	Computerized testing	Based on IRT models, established CAT item banks, item-selection strategies, ability estimation, termination strategies, and other technical foundations of testing	Item-level CAT testing forms, including the Bayes adaptive testing model and the maximum-likelihood adaptive testing model. (1) School academic ability tests (2) College and graduate school entrance tests
Current development stage of CAT (since 2007)	Item response theory	Computerized testing	Based on IRT models, and integrating cognitive diagnosis models and sequential probability ratio test principles, CD-CAT and CCT were developed; MST, i.e., testlet-level CAT, was developed from item-level CAT.	IRT-based CD-CAT; MST; CCT. Meanwhile, item-level CAT also exists. (1) Cognitive diagnostic tests (2) School academic ability tests (3) College and graduate school entrance tests (4) Professional qualification tests

Adaptive testing began as early as the Binet-Simon intelligence test in 1905. Through the ingenious use of mental age as an indicator, it placed the examinee's intelligence level and item difficulty level on the same scale, and was widely popularized and applied throughout the world. Later, researchers applied this form of adaptive testing to school academic ability testing, and continuously explored and improved the measurement theory, testing format, and measurement technology of adaptive testing. They proposed item response theory models and a system of theoretical assumptions for item response theory, and used IRT models to define examinee ability and item parameters on the same scale. Since 1975, CAT measurement technology has been based mainly on item response theory. A variety of measurement techniques and methods have been developed in areas such as CAT item-selection strategies, item exposure-rate control, ability estimation methods, termination strategies, and online calibration of CAT item parameters, and these methods began to be applied to large-scale tests such as college entrance examinations and graduate entrance examinations. Since 2007, CAT measurement technology has developed in three directions according to the needs of testing practice: IRT-based computerized adaptive testing for cognitive diagnosis, computerized multistage adaptive testing, and computerized classification testing, which are applied respectively to cognitive diagnosis in school teaching, college and graduate entrance examinations, and professional qualification tests.

References:

[1] Green B. Comments on tailored testing[C]. // Holtzman W. H. (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row. 1970.

[2] Weiss D J, Betz N E. Ability measurement: Conventional or adaptive? (Research Report 73-1)[R]. Minneapolis: University of Minnesota. (NTIS No. Ad 757788). 1973.

[3] Weiss D J. Adaptive testing by computer[J]. Journal of Consulting and Clinical Psychology, 1985, 53: 774-789.

[4] Killcross M C. A review of research in tailored testing(Report APRE No. 9/76)[R]. Franborough, Hants, England: Ministry of Defense, Army Personnel Research Establishment. 1976.

[5] Weiss D J. Adaptive testing by computer[J]. Journal of Consulting and Clinical Psychology, 1985, 53: 774-789.

[6] Meijer R R., & Nering M L. Computerized adaptive testing: overview and introduction[J]. Applied Psychological Measurement,1999, 23(3): 187-194.

[7] Chang H H. Psychometrics behind computerized adaptive testing[J]. Psychometrika, 2015, 80(1): 1-20.

[8] Zhang Huahua, Cheng Ying. The development and future prospects of computerized adaptive testing (CAT)[J]. Examinations Research, 2005, 1(1): 12-24.

[9] Zhang Qinghua, Yuan Yiping, Zhang Houcan. Development of adaptive testing: history and current status[J]. Psychological Exploration, 2006, 26(2): 84-87.

[10] Weiss D J, Sahin A. Computerized adaptive testing: From concept to implementation[M]. Guilford Press. 2024.

[11] Weiss D J, Yan D. A brief history of computerized adaptive and multistage testing[M]. // Yan D, von Davier, A A, Weiss D J. (Eds.). Research for Practical Issues and Solutions in Computerized Multistage Testing (1st ed.). Routledge. 2024.

[12] Moses T. A review of developments and applications in item analysis[M]. // Bennett R, von Davier M. (eds). Advancing human assessment. methodology of educational measurement and assessment. Springer, Cham. 2017.

[13] IACAT. First adaptive test–Binet' s IQ test[EB/OL]. 2025-10-11. Available online: http://iacat.org/first-adaptive-test

[14] Thurstone L L. A method of scaling psychological and educational tests[J]. Journal of Educational Psychology, 1925, 16: 433–451.

[15] Symonds P M. Choice of items for a test on the basis of difficulty[J]. Journal of Educational Psychology, 1929, 20(7): 481-493.

[16] Spache G. Serial testing with the revised stanford-binet scale, form 1, in the test range ii-xiv[J]. American Journal of Orthopsychiatry, 1942, 12(1): 81-86.

[17] Hutt M L. A clinical study of "consecutive" and "adaptive" testing with the revised Stanford-Binet[J]. Journal of Consulting Psychology, 1947, 11, 93-103.

[18] Ferguson G A. Item selection by the constant process[J]. Psychometrika, 1942, 7(1): 19-29.

[19] Lawley D N. On problems connected with item selection and test construction[J]. Proceedings of the Royal Society of Edinburgh, 1943, 61(3): 273-287.

[20] Lord F M. A theory of test scores[J]. Psychometric Monographs,1952, NO.7.

[21] Hick W E. Information theory and intelligence tests[J]. British Journal of Psychology, Statistical Section, 1951, 4: 157-164.

[22] Cowden D J. An application of sequential sampling to testing students[J]. Journal of the American Statistical Association, 1946, 4: 547-556.

[23] Angoff W H, Huddleston, E. M. The multi-level experiment: a study of a two-level test system for the college board scholastic aptitude test. Educational Testing Service Statistical Report SR-58-21.Princeton, New Jersey. 1958.

[24] Krathwohl D R, Huyser R J. The sequential item test (SIT)[J]. American Psychologist,1956, 2: 419.

[25] Bayroff A G. Feasibility of a programmed testing machine(Research Study No. 64-63)[R]. Washington, DC: U.S. Army Personnel Research Office. 1964.

[26] Moonan W J. Some empirical aspects of the sequential analysis technique as applied to an achievement examination[J]. Journal of Experimental Education, 1950, 18: 195-207.

[27] Lord F M. Some test theory for tailored testing[R]. ETS Research Bulletin Series, i-62. 1968.

[28] Lord F M. Individualized testing and item characteristic curve theory[R]. ETS Research Bulletin Series, i-37. 1972.

[29] Zhang Huahua, Wang Wenyi. "Internet+" assessment: the road to adaptive learning[J]. Journal of Jiangxi Normal University (Natural Science Edition), 2016, 40(5): 441-455.

[30] Lord F M. The 'ability' scale in item characteristic curve theory[J]. Psychometrika, 1975, 40(2): 205-217.

[31] Owen R J. A Bayesian approach to tailored testing[R]. ETS Research Bulletin Series, i-24. 1969.

[32] Owen R J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing[J]. Journal of the American Statistical Association, 1975, 70: 351-356.

[33] Lord F M, Novick M R. Statistical theories of mental test scores[M]. Boston, MA: Addison-Wesley. 1968.

[34] Wrigh B. Panchapakesan N. A procedure for sample-free item analysis[J]. Educational and Psychological Measurement, 1969, 29: 23-48.

[35] Weiss D J. Strategies of adaptive ability measurement (Research Report 74-5)[R]. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. 1974.

[36] Lord F M. Tailored testing, an application of stochastic approximation[J]. Journal of the American Statistical Association, 1971, 66: 707-711.

[37] Cleary T A, Linn R L, Rock D A. Reproduction of total test score through the use of sequential programmed tests[J]. Journal of Educational Measurement, 1968, 1: 183-187.

[38] Lord F M. The self-scoring flexilevel test[J]. Journal of Educational Measurement, 1971, 8: 147-151.

[39] Wood R. Response-contingent testing[J]. Review of Educational Research, 1973, 43(4): 529-544.

[40] Weiss D J. The stratified adaptive computerized ability test (Research Report 73-3)[R]. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. 1973.

[41] Wood R. The efficacy of tailored testing[J]. Educational Research, 1969, 11: 219-222.

[42] Novick M R. Bayesian methods in psychological testing[R]. Princeton, N. J.: Educational Testing Service, Research Bulletin RB-69-31. 1969.

[43] Urry V W. A monte-carlo investigation of logistic test models. Unpublished Doctoral Dissertation. Purdue University. 1970.

[44] Lord F M. Practical applications of item characteristic curve theory[J]. Journal of Educational Measurement, 1977, 14(2): 117-138.

[45] Weiss D J, Bock R D. New horizons in testing: Latent trait test theory and computerized adaptive testing[M]. New York: Academic Press. 1983.

[46] Clark C. Proceedings of the first conference on computerized adaptive testing (Washington, D. C., June 12-13, 1975)[C]. Superintendent of Documents, U.S. Government Printing Office. 1976.

[47] Lord F M. Applications of item response theory to practical test problems[C]. Paper presented at the convention of the American Psychological Association, Washington, D. C. 1976.

[48] Warm T A. A primer of item response theory (Technical Report 940279)[R]. Oklahoma City, Okla.: U.S. Coast Guard Institute. 1978.

[49] Lord F M. Applications of item response theory to practical testing problems[M]. Hillsdale, NJ: Lawrence Erlbaum. 1980.

[50] Hambleton R K, Swaminathan H, Rogers H J. Fundamentals of item response theory[M]. Newburry Park, CA: SAGE. 1991.

[51] van Der Linden W J. Constrained adaptive testing with shadow tests[M].// Van Der Linden W J, Glas C A. (Eds.), Computerized adaptive testing: Theory and Practice. Boston, MA: Kluwer Academic Publishers. 2000.

[52] Jian Xiaozhu, Dai Haiqi, Zhang Minqiang, Peng Chunmei. A classification overview of CAT item-selection strategies[J]. Psychological Exploration, 2014, 34(5): 446-451.

[53] Sympson J B, Hetter R D. Controlling item exposure rates in computerized adaptive testing[C]. Proceedings of the 27th annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development. 1985.

[54] Li Mingyong, Zhang Minqiang, Jian Xiaozhu. A review of test-security control methods in computerized adaptive testing[J]. Advances in Psychological Science, 2010, 18: 1339-1348.

[55] Lord F M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee' s ability[J]. Psychometrika, 1953, 18(1): 57-76.

[56] Bock R D, Mislevy J R. Adaptive EAP estimation of ability in a microcomputer environment[J]. Applied Psychological Measurement, 1982, 6: 431-444.

[57] Stocking M L. Scale drift in on-line calibration (Research Rep. 88-28)[R]. Princeton, NJ: ETS. 1988.

[58] Wainer H, Mislevy R J. Item response theory, item calibration, and proficiency estimation[M].// Wainer H, Dorans N J, Flaugher R, Green B F, Mislevy R J, Steinberg L, Thissen D. (Eds.). Computerized adaptive testing: A primer(Chap. 4, pp. 65–102). Hillsdale, NJ: Erlbaum. 1990.

[59] Wainer H, Dorans N J, Flaugher R, Green B F, Mislevy R J, Steinberg L, Thissen D. (Eds.). Computerized adaptive testing: A primer[M]. Hillsdale NJ: Erlbaum. 1990.

[60] Sands W A, Waters B K, McBride J R. Computerized adaptive testing: From inquiry to operation[M]. Washington (DC), American Psychological Association.1997.

[61] Drasgow F, Olson-Buchanan J B. (Eds.). Innovations in computerized assessment[M]. Hillsdale NJ: Erlbaum. 1999.

[62] Van Der Linden W J, Glas C A. Computerized adaptive testing: Theory and Practice[M]. Boston, MA: Kluwer Academic Publishers. 2000.

[63] Luo Shuqing. Applications of modern measurement theory in examinations[M]. Wuhan: Central China Normal University Press. 2003.

[64] Jian Xiaozhu, Zhang Minqiang. The concept, types, and characteristics of computerized adaptive testing based on IRT[J]. China Examinations, 2024, 31(9): 66-75.

[65] Yan D, von Davier A A, Weiss D J. (Eds.). Research for Practical Issues and Solutions in Computerized Multistage Testing [M]. 1st ed. Routledge. 2024.

[66] Weiss D J., Sahin A. Computerized adaptive testing: From concept to implementation[M]. Guilford Press. 2024.

[67] Chen Ping, Ding Shuliang, Lin Haiqing, Zhou Jie. Item-selection strategies for computerized adaptive testing under the graded response model[J]. Acta Psychologica Sinica, 2006, 38(3), 461-467.

[68] Dai Haiqi, Chen Dezhi, Ding Shuliang, Deng Tai 萍. A comparison of item-selection strategies for computerized adaptive testing with polytomously scored items[J]. Acta Psychologica Sinica, 2006, 38(5), 778-783.

[69] Cheng Xiaoyang, Ding Shuliang, Yan Shenhai, Zhu Longyin. Item-selection strategies for computerized adaptive testing introducing an exposure factor[J]. Acta Psychologica Sinica, 2011, 43: 203-212.

[70] Zhu Longyin, Ding Shuliang, Wang Qianjuan. A study on classification-accuracy stratified stopping rules for variable-length CAT[J]. Psychological Exploration, 2008, 28(4): 80-84.

[71] Chen Ping, Ding Shuliang. Computerized adaptive testing allowing review and answer changes[J]. Acta Psychologica Sinica, 2008, 40(6): 737-747.

[72] Chen Ping. Two new online calibration methods for computerized adaptive testing[J]. Acta Psychologica Sinica, 2016, 48(9): 1184–1198.

[73] Liu Yan, Dai Jing, Shi Xiaolian, Niu Yu, Zhu Jiating, Gu Xiaoqing. Application and implications of cognitive diagnostic theory in computerized adaptive testing[J]. Distance Education in China, 2017, 37(4): 42-49.

[74] Yang Tao, Huang Shengshun, Xin Tao. The basic structure of computerized multistage testing and its research progress[J]. China Examinations, 2015, (07): 7-12.

[75] Wang Yanjun, Luo Zhaosheng, Wang Rui. A review of research on computerized multistage adaptive testing[J]. Psychological Science, 2015, 38(02): 452-456.

[76] Jian Xiaozhu, Chen Ping. A review of the characteristics and development of computerized classification testing[J]. Examination Research, 2020, (06): 77-89.

[77] Ren He, Huang Yingshi, Chen Ping. Categories, characteristics, and applications of stopping rules in computerized classification testing[J]. Advances in Psychological Science, 2022, 30(5), 1168-1182.

[78] Chai Shengsan. A study on the remote CAT reading-test model for the Chinese Proficiency Test (HSK)[J]. Distance Education in China, 2013, 45(11): 81-87.

[79] Tian Jianquan, Miao Danmin, Yang Yebing, et al. Development of an adaptive computerized jigsaw-puzzle test for conscripted citizens[J]. Acta Psychologica Sinica, 2009, 41(2): 167-174.

[80] Wang Lei. An exploration of the application of artificial-intelligence generated content technology in educational examinations[J]. China Examinations, 2023, 30(8): 19-27.

[81] Chen Zhe, Lang Weimin, An Haiyan, et al. A proposition-intelligence assistance system based on large language models[J]. Telecommunications Express, 2024, (4): 1–6.

[82] Han Yuting, Wang Wenxuan, Liu Hongyun, et al. Technological innovation and real-world challenges in automatic item generation[J]. Advances in Psychological Science, 2025, 33(10): 1766–1782.

[83] Chen P. Reinforcement Learning for Item Selection in CAT[J]. Psychometrika, 2022, 87(3): 932–958.

[84] Wang P, Liu H, Xu M. An adaptive testing item selection strategy via a deep reinforcement learning approach[J]. Behavior Research Methods, 2024, 56(8): 1–20.

[85] Cardwell R, Naismith B, LaFlair G T, Nydick S. Duolingo English test: Technical manual (Duolingo Research Report)[R]. Duolingo. 2024.

[86] Burstein J, LaFlair G T, Kunnan A J, von Davier A A. A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test (Duolingo Research Report DRR-22-01)[R]. Duolingo. 2022.

[87] Langenfeld T, Burstein J, von Davier A A. Digital-first learning and assessment systems for the 21st century[J]. Frontiers in Education, 2022, 7, 857604. https://doi.org/10.3389/feduc.2022.857604

Review of the Development of Measurement Theory and Technology for Computerized Adaptive Test

Xiaozhu Jian [1]

Department of Psychology, School of Public Policy and Management, School of Economics and Management,
Nanchang University; Nanchang, Jiangxi, 330036, China

Abstract:

The conceptual origins of computerized adaptive testing (CAT) can be traced back to the Binet-Simon Intelligence Test in 1905. In the late 1960s, based on item response theory, the measurement theory and technology of CAT achieved breakthroughs and developed further. This paper proposes a five-stage developmental framework for CAT, grounded in the key clue that item difficulty parameters and examinee ability parameters are estimated and built on a common metric scale. The stages are delineated according to key evolutionary milestones in CAT: the embryonic stage of adaptive testing, the early development stage, the breakthrough stage, the stage of refinement, and the ongoing development stage. For each stage, the paper elaborates on the theory and technology, highlights salient features, and discusses practical applications of CAT. Furthermore, the paper summarizes the latest cutting-edge research prospects of AI technology applied to CAT.

Key Words: Adaptive Testing, Computerized Adaptive Testing, item response theory, Measurement theory

Note: Figure translations are in progress. See original paper for figures.