Journal of Classification 30 (2013)
DOI: 10.1007/s00357-013-9128-5

The Development of Computerized Adaptive Testing with Cognitive Diagnosis for an English Achievement Test in China

Hong-Yun Liu

Beijing Normal University, China

Xiao-Feng You

Foreign Language Teaching and Research Press, China

Wen-Yi Wang and Shu-Liang Ding

Jiangxi Normal University, China

Hua-Hua Chang

University of Illinois at Urbana-Champaign, U.S.A.

The authors are indebted to the editor, associate editor and three anonymous reviewers for their constructive suggestions and comments on the earlier manuscript. They particularly thank Carolyn Anderson for her suggestions which led to numerous improvements.

This research was partially supported by Natural Science Foundation of China Grant 31100759, 30860084, and 31160203, by the project of Student Academic Quality Evaluation under the National Centre of Curriculum and Textbook Development for Basic Education (NCCT), Ministry of Education of the People's Republic of China. Part of the paper was originally presented at the 2010 Annual Meeting of the National Council on Measurement in Education, Denver Co.

Corresponding Author's Address: Hong-Yun Liu, Beijing Normal University, 19 Xinjiekou Wai Street, Beijing 100875, P. R. CHINA, e-mail: hyliu@bnu.edu.cn.

Abstract: Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) is an innovative development in psychological and educational measurement. The CD-CAT is designed to select items sequentially during the process of assessment to provide accurate information about examinees' cognitive strengths and weaknesses. This paper introduces the development and implementation of a web-based CD-CAT program for a large-scale English test in China, the Level 2 English Achievement. The paper is organized as follows. First, the main components of the CD-CAT design will be introduced, such as identification of attributes, Q-matrix formation, choice of CD model, item bank and calibration, item selection strategy, the parameter estimation, and design of CAT delivery system. Next, a study investigating an adequacy of model fitting for a large-scale pretest is described. This is followed by a classification accuracy study using the item parameters calibrated from the pretest. An ensuing section presents the results of a field CD-CAT test with 582 students in Beijing. A validity study of the computer-generated diagnostic results was also conducted. The final section discusses directions for future work.

Keywords: Cognitive Diagnostic Assessment (CDA); Computerized Adaptive Testing (CAT); Diagnostic feedback; DINA model; English Level-2 Test; Validity.

1. Introduction

Cognitive diagnostic assessment has become a promising method that offers informative feedback for each examinee rather than simply offering a summative total score or subscale scores. Diagnostic feedback on attributes and skills specified to account for the students’ performance provides the content areas in which remedial instruction is needed (Leighton, Gierl and Hunka 2004). Therefore, it can be not only used to evaluate students’ achievement, but to provide detailed and valuable information on each student’s strengths and weaknesses in learning.

The research on cognitive diagnostic models (CDMs) over the past three decades has essentially focused on modeling and calibrating items. Recently, many of the models have been successfully applied in real applications, including the application of the Rule Space Model (RSM) for fraction addition in SAT testing data (Tatsuoka 1995), the Bayesian probability inference model for fraction subtraction (Mislevy 1994), the General Diagnosis Model (GDM) for NAEP data (Xu and von Davier 2006), and the development of software for the Fusion Model (Templin 2005) and its diagnostic application recently reported by the ETS (Educational Testing Service 2004).

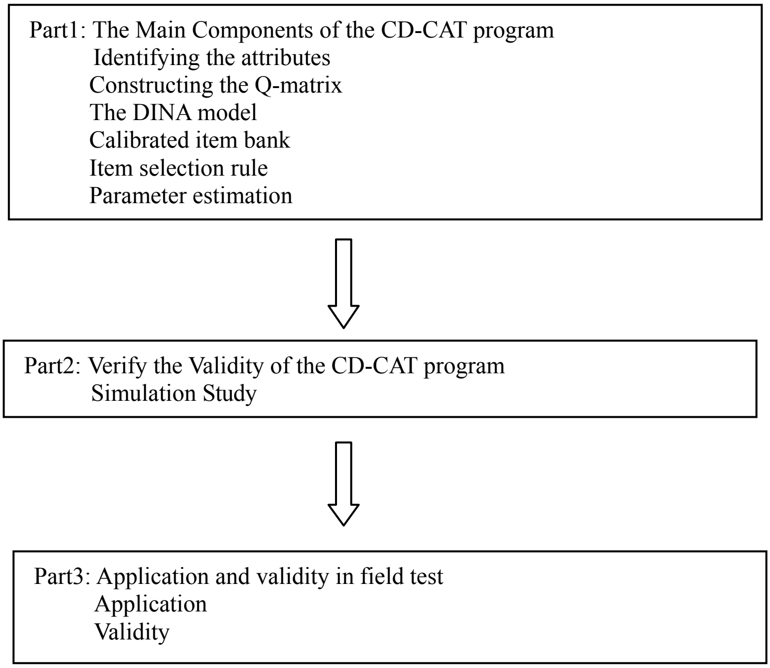
Recently, much research has focused on methods and issues to facilitate practical applications of cognitive diagnostic models. Several theories and algorithms were developed to implement cognitive diagnostic computerized adaptive assessments (Xu, Chang, and Douglas 2003, 2005; McGlohen 2004; Cheng and Chang 2007; McGlohen and Chang 2008; and Cheng 2009) that are referred to as CD-CAT. Developing CAT as a diagnostic tool for assessment and evaluation has always been a primary interest. Tatsuoka (1997) developed a CAT based on RSM and showed its potential application for fraction addition. Jang (2008) described the possible utility of CD-CAT in a classroom setting: teachers could use CD-CAT to diagnose specific skills or knowledge taught in each unit. Students can also immediately receive the reports about their strengths and weaknesses after completing the exam on the computers.

To develop an item selection algorithm for CD, a flexible cognitive diagnostic model needs to be chosen. Throughout the progression of cognitive diagnosis research, an abundance of models have been proposed to provide cognitively diagnostic information in the assessment process (see Hartz, Roussos, and Stout 2002). What distinguishes models from one another are the assumptions that dictate how attributes are utilized to construct responses. Among these models, the Deterministic Inputs, Noisy "And" Gate (DINA) model (Macready and Dayton 1977, Junker and Sijtsma 2001) has been used widely by researchers and practitioners in simulation studies and large scale implementations. It features the simplicity of estimation and interpretation. Based on the likelihood function and item responses, both item parameters and examinees' cognitive profiles can be conveniently estimated using maximum likelihood estimation (MLE). The DINA model has been studied extensively and many findings are encouraging. For example, de la Torre (2008), Cheng (2009), and de la Torre and Lee (2010) reported that the diagnoses based on DINA model are accurate when the Q-matrix is correctly constructed.

The purpose of the current project is to develop a CD-CAT for large-scale application. Our main objective is to develop an on-line assessment system to combine CAT with CD and provide cognitive diagnostic feedback to the test-takers of the Level 2 English Achievement in China. The paper is organized as follows. First, the main components of the CD-CAT design will be introduced, such as identification of attributes, Q-matrix formation, choice of CD model, item bank and calibration, item selection strategy, the parameter estimation, and design of CAT delivery system. Subsequently, a Monte Carlo simulation study to evaluate the system design is described. An ensuing section presents the results of a CD-CAT field test administered to 582 students in Beijing that is followed by a validity study. The final section discusses directions for future work. The major topics mentioned in the present paper are summarized in Figure 1.

2. Components of the CD-CAT System

In China, the English language proficiency for compulsory education is divided into six levels by the *National English Curriculum Standard*s. Among them, Level 2 is set for Grade 6 students. The objective



Part1: The Main Components of the CD-CAT program
Identifying the attributes
Constructing the Q-matrix
The DINA model
Calibrated item bank
Item selection rule
Parameter estimation

Part2: Verify the Validity of the CD-CAT program
Simulation Study

Part3: Application and validity in field test
Application
Validity

Figure 1. The framework of the main topics covered by the paper.

of the project is to develop a cognitive diagnostic assessment not only to help students to meet the standards, but to assist teachers to instruct more efficiently. The main components of the system design are described below.

2.1 Identifying Attributes and Constructing the Q-Matrix

General attributes and Q-matrices can be identified by content experts after items and tests are developed. Such a post-hoc approach of constructing a Q-matrix may not be ideal since the attributes can only be matched with the items currently available in the item pool. As a result, an attribute being assessed may not have a sufficient number of items to measure it. A different approach is to make the content experts define the attributes and relationships among them before test development. Under this alternative approach, the test developers can write items following instructions about the attributes, relationships, and other pre-specified requirements. In the present study, all the attributes and relationships were defined before test development, and the test developers wrote items following guidelines on the attributes and their relationships.

One of the most important steps in the development of an attribute-based assessment is defining and constructing the attributes. In our study, eleven content experts were invited and they identified eight attributes, for this purpose. In addition, an adjacency matrix, the reachability matrix and the student matrix Qs whose columns represent all kinds of the knowledge stases, is also constructed. Note that Qs here is the same as the Tatsuoka's matrix Qr with one more zero vector. Based on the matrix Qs or a sub-matrix of Qs, the test items can be designed. According to Leighton and Gierl (2007), The Qs has a particularly important interpretation: it represents the cognitive blueprint for the test. The sub-matrix of Qs, called Qt, now can be viewed as a test cognitive blueprint that provides useful information for the item writers. Based on the attributes and test blueprint, the 11 content experts wrote 400 multiple-choice items, including Listening Comprehension, Grammar & Vocabulary, and Reading Comprehension.

For quality control, each item was tested by eight Grade 6 students selected from different English proficiency levels. During the pilot testing, each student was asked to report what skills and strategies he/she used to answer the item and why. The outcomes of the think-aloud protocol were then analyzed by the researchers and content experts, and some definitions of the attributes were subsequently revised. After a while, eight attributes (see Appendix 1) were identified. Note that, as an initial study, these attributes were treated independently.

Like most cognitive diagnostic assessments, the implementation of the CD-CAT requires construction of a Q-matrix (Tatsuoka 1983) to describe the relationship between the items and the attributes. In our study, researchers and content experts constructed the Q-matrix according to the test cognitive blueprint. In the Q-matrix, each of the 400 items is listed in a separate row, and each of 8 attributes in a separate column. Then, six English teachers were recruited to evaluate the reasonableness of the Q-matrix structure. They were asked to evaluate whether the attributes defined in the Q-matrix were needed to correctly answer each item, and write down any potential attributes that were not included on the list. If there was disagreement, discussion took place. If the discussion failed to reach an agreement, the item was deleted.

A Paper-and-Pencil based pretest was administered to more than 38,600 students from 78 schools in 12 counties. Based on the students' responses, the parameters of the 3-parameter logistic model (3PLM) were estimated. To verify the construct validity of the Q-matrix, a regression analysis (Yang and Embretson 2007) of item difficulty (3PLM) on the attribute variables for each item was conducted. Let the eight attributes be denoted by A_1, A_2, … , and A_8, where $A_k = 1$ if the attribute k is required to correctly answer the item, and $A_k = 0$, otherwise. The regression study indicated that all the attribute variables have high impact on the item difficulties (all have R^2 around 0.455). Both standardized and unstandardized regression coefficients are presented in Table 1. Most attributes showed significant effects in predicting item difficulty. Note that the Beta parameters (standardized regression coefficient) of each attribute can be interpreted as an indicator of its relative weight in predicting the item difficulty. These results are consistent with the findings of Yang and Embretson (2007).

2.2 The DINA Model

The purpose of cognitive diagnostic analysis is to identify which attributes are mastered by the examinees. For each examinee, the mastery profile will be translated into a vector: $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{ik})'$, where $\alpha_{ik} = 1$ indicates that the *i*th examinee masters the *k*th attribute and $\alpha_{ik} = 0$ otherwise. The DINA model described in the following equation is employed in the CD-CAT system taking advantage of its simplicity of estimation and interpretation:

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i) = g_j^{1-\eta_{ij}} (1 - s_j)^{\eta_{ij}}, \quad (1)$$

where $\boldsymbol{\alpha}_i$ represents the column vector of knowledge state for examinee *i* with components of α_{ik}, which equals either 0 or 1, and $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, where *K* is the number of attributes, and q_{jk} indicates whether skill *k* is required to correctly answer item *j*. If examinee *i* has mastered all the attributes required for item *j,* then $\eta_{ij} = 1$, otherwise, $\eta_{ij} = 0$. A ‘slipping’ parameter s_j is defined as $s_j = P(X_{ij} = 0 | \eta_{ij} = 1)$, the probability of an incorrect response on item *j* when an examinee has mastered all the attributes. A ‘guessing’ parameter g_j equals $g_j = P(X_{ij} = 1 \mid \eta_{ij} = 0)$ and refers to the probability that an examinee who has not mastered all the attributes of *j*th item but answers correctly. Under the assumption of local independence, the joint likelihood function of the DINA model $L(s, g; \alpha)$ can be easily obtained (e.g., de la Torre 2009).

Table 1. Linear Regression Coefficients of Attributes for Difficulty

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	b	Std. Error	Beta		
(Constant)	-1.469	.506		-2.906	.007
A1	.173	.405	.090	.427	.673
A2	1.339	.502	.579	2.669	.013
A3	1.198	.560	.623	2.138	.042
A4	.880	.577	.410	1.525	.139
A5	1.787	.550	.702	3.248	.003
A6	1.696	.699	.486	2.427	.022
A7	.728	.540	.315	1.349	.188
A8	1.388	.495	.480	2.804	.009

2.3 Item Bank and Calibration

For the purpose of booklet linking, thirteen booklets were assembled. One booklet consists of 40 anchor items served as a central linking booklet in our project, and the other twelve booklets included 30 new items and 10 anchor items. More than 38,600 students from 78 schools in 12 counties took the pretest. In each school, 30 students were randomly sampled to take the anchor test, while the other students were randomly divided into the 12 groups and were assigned one booklet each. Based on the students' responses, the parameters of DINA model and 3PLM were estimated. The distribution of the parameters across the item bank is summarized in Table 2. In addition, items that assessed a single attribute and items that assessed more than one attribute were included in the item bank. There were 330 items that assessed a single attribute, and 22 items that assessed two attributes. The average number of attributes was 1.062 per item. The number of items assessing attributes A_1 through A_8 equaled with 57, 45, 80, 60, 40, 20, 40, 32, respectively, with an average of 47 items per each attribute.

2.4 Item Selection Strategy

The most important component in CD-CAT is the item selection rules. In our study, a set of eight items was randomly assigned to each examinee, one from each attribute category at the beginning of the testing. The test becomes adaptive after the student completes the first 8 items. The early studies have shown that a procedure based on Shannon Entropy (SHE) performs well in terms of classification rates (Tatsuoka and Fergu-

Table 2. The Distribution of Item Parameters of IRT Model and DINA Model Across the Item Bank.

	IRT		DINA		
	a	b	c	s	g
Mean	0.864	-0.018	0.132	0.222	0.305
SD	0.374	0.550	0.040	0.100	0.074
Min	0.300	-1.511	0.009	0.004	0.083
Max	2.005	1.527	0.292	0.399	0.400

son 2003; Xu and Douglas 2006; Xu, Chang, and Douglas 2003). Therefore, the SHE procedure was employed in the CD-CAT program.

The SHE for a discrete random variable *X* (i.e., an item) is defined as

$$H(X) = -\sum_{X} P(x)\log P(x), \tag{2}$$

where $0\log 0 = 0$. Shannon Entropy measures the uncertainty associated with the distribution of a random variable. Intuitively, the item that produces the smallest expected value of SHE is associated with the least amount of uncertainty in the test taker’s attribute pattern distribution and therefore, will be chosen as the next item. In our design, the posterior distribution of the knowledge state is computed with the SHE item selection method. After *t* items, the posterior distribution of the knowledge state is computed as f_t (f_0 could be set as uniform distribution). The SHE of the posterior distribution f_t is

$$H(f_t) = -\sum_{c=0}^{2^K-1} f_t(\alpha_c)\log f_t(\alpha_c). \tag{3}$$

For the *(t+1)*th item, the probability of observing *x* is

$$\Pr(X_{ij} = x \mid u_i^{(t)}) = \sum_{c=0}^{2^K-1} P(X_{ij} = x \mid \alpha_c) f_t(\alpha_c), \tag{4}$$

where f_t is a prior distribution, and X_{ij} is item response for examinee *i* and item *j*. The conditional posterior distribution of the knowledge state becomes

$$f_{t+1}(\alpha_c \mid X_{ij} = x) = \frac{f_{0c} L(\alpha_c; u_i^{(t)}) P_j(\alpha_c)^x (1 - P_j(\alpha_c))^{1-x}}{\phi_{it+1}}. \tag{5}$$

Here $L(\alpha_c; u_i^{(t)}) = \prod_{j=1}^{t} P_j(\alpha_c)^{u_{ij}} (1 - P_j(\alpha_c))^{1-u_{ij}}$ is the likelihood function of $u_i^{(t)}$, $P_j(\alpha_c)$ can be computed from formula (1), and $\varphi_{t+1} = \sum_{c=0}^{2^K-1} f_t(\alpha_c) P_j(\alpha_c)^x (1-(\alpha_c))^{1-x}$, where f_0 (α_c) is initial prior probability of the knowledge state α_c. Given $X_{ij} = x$, according to (3), the conditional entropy of f_{t+1} is just SHE of $f_{t+1}(\alpha_c \mid X_{ij} = x)$, denoted as $H(f_{t+1}(\alpha_c \mid X_{ij} = x))$. From formulas (4) and (5), the expectation of f_{t+1} conditional on random variable X is

$$SHE_{ij}(f_{t+1}) = \sum_{x=0}^{1} H(f_{t+1}(\alpha_c \mid X_{ij} = x)) \Pr(X_{ij} = x \mid u_i^{(t)}) . \quad (6)$$

Here, $SHE_{ij}(f_{t+1})$ refers to the expected entropy of examinee *i* on item *j* after *t* items are completed. The next item is selected to minimize SHE in Equation (6) from the remaining items in the bank after examinee *i* completed *t* items. See Xu, Chang, and Douglas (2003) for more details about the Shannon Entropy method.

Under the DINA model, one only needs to substitute $P_j(\alpha_c)$ (see formula (4) and formula (5)) with the item response function defined in (1), then SHE can be easily obtained. Although the purpose of using SHE is to select the next item based on examinee's previous performance so that the uncertainty of the posterior distribution for the knowledge state can be greatly reduced, SHE may lead to decreasing uncertainty of its posterior distribution after selecting the next item. In this way, selecting the item with the largest difference between (3) and (6) can also be taken as a strategy of selecting items, and it will terminate when the difference between (3) and (6) is below a certain predetermined threshold, but this is left for future research.

2.5 Examinee's Profile Estimation

The aim of CD-CAT is to classify an examinee's latent class into a multidimensional latent vector with binary entries. The Maximum a Posterior Estimation (MAP) method was used in the program. When the prior distribution is uniform, MAP is equivalent to maximum likelihood estimation. The MAP is used to find the knowledge state that would maximize

$$\hat{\alpha}_i = \arg \max_{\alpha_c = 0,1,\cdots,2^K-1} (P(\alpha_c \mid \underset{i}{u}^{(m)})), \tag{7}$$

where $P(\alpha_c \mid \underset{i}{u}^{(m)})$ is a posterior probability and it can be calculated from (5). In the current study, we only consider fixed length CAT and let *m* be the test length. From the posterior probability $P(\alpha_c \mid \underset{i}{u}^{(m)})$, the sum of each attribute margin is calculated. For example, the marginal probability of the attribute *k* (marginal posterior probability method, MPPE) equals

$$mp_{ki} = \sum_{c=0}^{2^K-1} P(\alpha_{kc} \mid \underset{i}{u}^{(m)})\alpha_{kc}. \tag{8}$$

MPPE can be used to calculate the probability that attribute *k* is mastered. Also, it can be used to derive the knowledge states of mastery or non-mastery for the attribute according to the cut point. Here, $\alpha_{ki} = 1$, if $mp_{ki} > cutpoint$, and otherwise, $\alpha_{ki} = 0$.

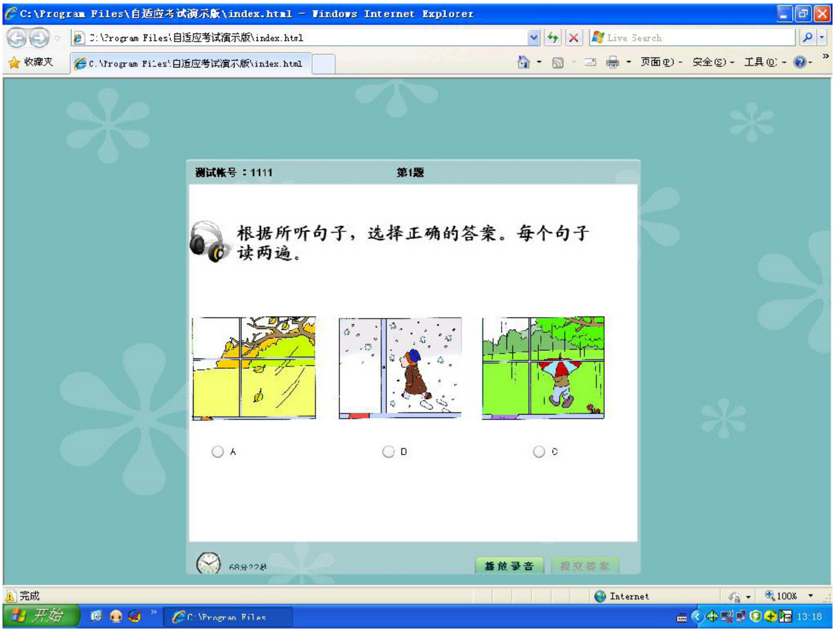
2.6 Design of CAT Delivery System

The cutting-edge Browser/Server (B/S) architecture allows the school to implement CAT with little additional cost because it uses its current computer labs and networks. The B/S architecture uses commonly available web-browsing software on the client side and a simple server. The CAT server can be installed on a laptop that is connected to the school's existing network of PCs and Macs. To make CAT diagnostic tools available in many schools, the system design has taken advantage of a browser-based test delivery application. As a result, schools and districts will be able to make use of their existing computers and network equipment. Figure 2 shows the system uses IE 9 to deliver the CD-CAT. Note that all the browser options shown in the upper frame can be easily blocked so that test takers have no access to test related information.

3. Simulation Studies

3.1 Simulation Design and Data Generation of CD-CAT

Monte Carlo simulation studies were conducted with the pre-calibrated item parameters to validate the functions of the CD-CAT system, including the item selection algorithm, estimate accuracy, and classification consistency. Three different test lengths, 20 items, 30 items and 36 items, were employed. For each condition, 1,000 simulated examinees were generated uniformly from the space of possible attribute

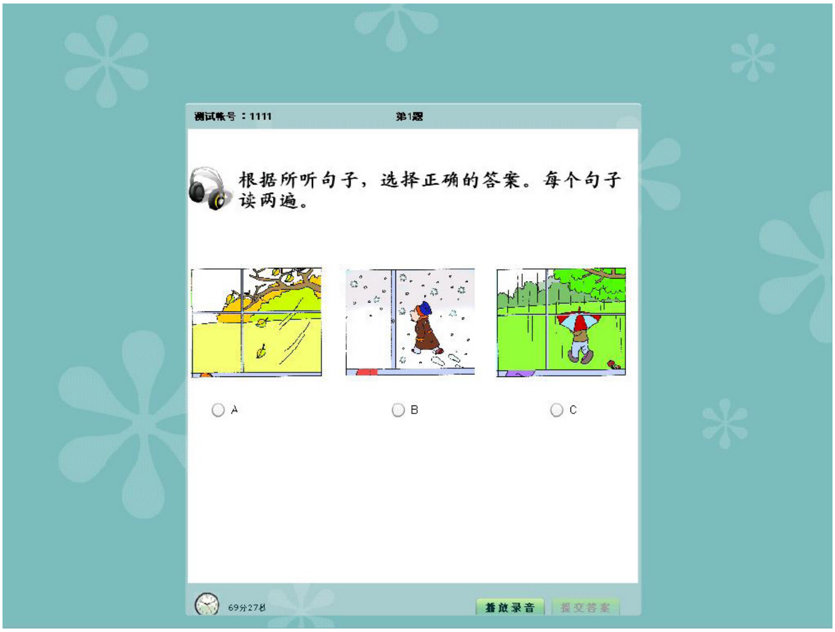

C:\Program Files\自适应考试演示版\index.html - Windows Internet Explorer
C:\Program Files\自适应考试演示版\index.html
Live Search
收藏夹
C:\Program Files\自适应考试演示版\index.html
页面(P)
安全(S)
工具(O)
播放录音
提交答案
完成
Internet
100%
开始
C:\Program Files
13:18

测试帐号：1111　第1题

根据所听句子，选择正确的答案。每个句子读两遍。

○ A　○ D　○ C

68分22秒



测试帐号：1111　第1题

根据所听句子，选择正确的答案。每个句子读两遍。

○ A　○ B　○ C

69分27秒

Figure 2. The Browser/Server (B/S) architecture allows Web browser (IE 9 here) to conveniently deliver a multimedia rich individualized assessment to any PC that is connected to the Internet. Note that all the web options shown in the upper frame can be easily blocked so that test takers have no access to test related information.

patterns. Both content constrained and non-content constrained situations were investigated in the simulation studies. The unconstrained item selection strategy selects the next item from all the remaining items in the item bank; whereas, the constrained strategy selects items according to the content specification (i.e., Listening Comprehension, Grammar and Vocabulary, and Reading Comprehension).

In the simulation study, three estimation methods were applied to classify the knowledge states: (1) the Method A (Leighton 2004), (2) the Log Likelihood ratio method (LL method; Zhu and Ding 2008), and (3) Maximum Likelihood estimation method (MLE; de la Torre 2009). The examinees' abilities were estimated by the EAP method (Bock 1982). For each condition each of the three methods was repeated 30 times, and their performances were compared in terms of the average test information, rate of marginal match (RMM), and rate of pattern match (RPM).

RMM and RPM. Suppose there are *K* attributes in the cognitive diagnostic test. The set of 'true' knowledge states and the set of estimated knowledge states are respectively denoted as True-state and Estimate-state: True-state $=\{ \alpha_i \mid \alpha_i$ is the true knowledge state of examinee $i \}$, Estimate-state$=\{ \hat{\alpha}_i \mid \hat{\alpha}_i$ is the estimated knowledge state of examinee $i \}$. If $\alpha_i = \hat{\alpha}_i$, let $h_i = 1$, otherwise, $h_i = 0$ and let *RPM* be

$$RPM = \frac{1}{M} \sum_{i=1}^{M} h_i \,, \tag{9}$$

where *M* is the number of examinees. In other words, *RPM* is the average number of examinees whose knowledge state is estimated correctly.

The knowledge state is a latent vector, and the 'true' knowledge state is unknown. In a Monte Carlo simulation, however, the 'true' knowledge state is known and how well methods perform can be assessed. If $\alpha_{ij} = \hat{\alpha}_{ij}$, let $h_{ij} = 1$, otherwise, $h_{ij} = 0$ and let *RMM* be

$$RMM = \frac{1}{MK} \sum_{i=1}^{M} \sum_{j=1}^{K} h_{ij} \,. \tag{10}$$

In the CD-CAT simulation studies, examinees' response data were generated as follows: $P_j(\alpha_i)$ was calculated based on the item bank parameters using equation (1); a random number *r* was generated from the uniform distribution on (0, 1); and if $P_j(\alpha_i) > r$, examinee *i* scored 1 on the item *j*, otherwise 0.

Table 3. Simulation Results of Non-Content Constraint and Content Constraint

	Test Length	Estimation Method	Classification rate (%)		Test Information
			RPM	RMM	
Non-Content Constraint	20	A	63.88	92.61	9.5870
		LL	48.24	89.17	
		ML	80.17	93.92	
	30	A	71.24	94.49	13.9221
		LL	53.78	90.40	
		ML	90.03	95.78	
	36	A	75.55	95.43	16.1941
		LL	57.32	91.33	
		ML	92.47	96.20	
Content Constraint	36	A	39.62	88.52	
		LL	33.43	86.68	17.2480
		ML	69.37	95.49	

3.2 Simulation Results

Table 3 presents the correct classification rates of each attribute and the average test information for different test lengths, estimation methods, and content constraint status.

Table 3 shows that the content constrained and non-content constrained methods perform similarly regarding to *RMM*, but for the non-content method *RPM* is higher. The MLE method of the DINA model is significantly outperformed the other two. The average information is greater than 16 when the test length is fixed at 36 items for both the content constrained and non-content constrained methods, which may indicate that imposing content constraints do not sacrifice estimation efficiency

4. Field Test and Validity Study

In the real testing, the content constrained SHE procedure was employed to select items. The test was a 40-minute, 36-item, fixed-length web-based CAT.

4.1 Application of the CD-CAT Program

Participants: A total of 584 students in Grade 5 and Grade 6 from 8 schools in Beijing, 300 boys and 284 girls, participated in the CD-CAT field test. The descriptive statistics of the items, raw scores, and attributes are shown in Table 4. The test length is fixed at 36. Though the test is

Table 4. Descriptive Results of Number of Items and Raw Score for Each Attribute

	Number of Items				Number correct			
	Mean	SD	Min	Max	Mean	SD	Min	Max
A1	9.49	1.888	2	16	8.29	1.897	0	14
A2	6.66	2.004	1	12	5.79	2.037	0	10
A3	5.47	2.303	1	12	4.56	2.277	0	10
A4	4.78	1.017	2	8	4.54	.985	0	8
A5	3.65	1.094	1	7	2.84	1.272	0	5
A6	3.72	.743	2	7	2.94	1.174	0	6
A7	4.86	.960	3	11	3.91	1.487	0	11
A8	3.50	1.008	2	8	2.30	1.214	0	8

designed to get diagnostic information for the 8 attributes, for each student the average number of items measuring per attribute differed over the eight attributes. For example, each student would have about 9.49 items measuring Attribute 1, whereas 3.5 for Attribute 8.

Diagnostic Results: The skill profile of each participant was determined by classifying each skill into mastery or non-mastery states. The classification results are presented in Tables 5 and 6. Table 5 indicates that the mastery proportions for all the attributes are great than 60%. The results in Table 6 indicate that 37.16% of examinees master all of the 8 attributes, and 56% of examinees master 7 or 8 attributes. It can be concluded that most examinees have exceeded the minimum pass level of student achievement measured by the English Level 2 test.

4.2 Validity Study in the Field Test

A validity study was conducted to investigate whether the cognitive diagnostic results generated by the CD-CAT system are consistent with those obtained from an academic achievement test the students took earlier. Ninety students from three schools in Beijing were selected for the pilot validity study. Note that the academic achievement test reports students' performance levels ranging from excellent, good, basic, and below basic. As to the CD-CAT assessment, the number of attributes mastered for each examinee was reported. Table 7 summarizes the consistency between the students' classification results from CD-CAT and those from the achievement test. It is interesting to notice that 23 out of 27 participants who were classified as mastering all the 8 attributes in CD-CAT are reported as excellent in the academic achievement test.

Table 5. Proportions of Masters for Eight Skills for CD-CAT System

	Masters		Non-Masters	
Attributes	Number	Percentage	Number	Percentage
A1	564	96.58	20	3.42
A2	462	79.11	122	20.89
A3	434	74.32	150	25.68
A4	560	95.89	24	4.11
A5	406	69.52	178	30.48
A6	460	78.77	124	21.23
A7	388	66.44	196	33.56
A8	366	62.67	218	37.33

Table 6. The Distribution of the Number of Mastered Skills

Number of mastered attributes	Number of examinees	Percentage of examinees
0	3	0.51
1	8	1.37
2	22	3.77
3	38	6.51
4	47	8.05
5	55	9.42
6	83	14.21
7	111	19.01
8	217	37.16

Table7. The Consistency of Performance Levels with the Number of Mastered Attributes

	The number of mastered attributes									
Academic performance level	0	1	2	3	4	5	6	7	8	Total
Excellent	0	0	1	1	1	3	4	6	23	39
Good	0	0	1	2	8	5	7	7	3	33
Basic	1	1	3	5	3	1	0	0	1	15
Below Basic	0	1	2	0	0	0	0	0	0	3
Total	1	2	7	8	12	9	11	13	27	90

A correspondence analysis (similar to Doey and Kurta 2011) was conducted to examine the association between the classification results from CD-CAT and those from the academic test. The results indicated that the total variation can be explained up to 98.1% by three dimensions: 54.1%, 28.3%, and 15.7% for the first, second, and third dimensions respectively. The correlation between rows and columns on the first dimension is 0.736. Note that the chi square equals 88.287 (df=24, P<0.001), which clearly implies that the category variation of the academic test is highly consistent with that of CD-CAT. More specifically, students were classified as excellent by the academic test are tend to have mastered 7 or 8 attributes in the CD-CAT test, while 4 to 6 for those classified as Good, and 0 to 3 for those classified as "Below Basic Level".

5. Conclusions

The purpose of this paper was to introduce and illustrate a CD-CAT system that has a great potential for large-scale applications. The proposed cognitive diagnostic assessment has been shown to be effective in providing formative diagnostic feedback through a fine-grained reporting of learners' skill mastery profiles. The CAT part of the design has the ability to tailor a test to each examinee based on his/her cognitive latent class and the CD model has demonstrated great potential to be effectively used to inform learners of their cognitive strengths and weaknesses in the assessed skills. The CD-CAT system can provide the students, parents, and teachers with diagnostic reports that are downloadable from the website providing diagnostic information to promote instructional improvement. Over the past ten years in China, student assessment has become an increasingly important feature of the public education. Combining available benefits from both CD and CAT, the CD-CAT design is an example for the next-generation assessment that targets the needs of both accountability and instructional improvement.

In this paper, an array of methods concerning how to build an operational CD-CAT was introduced. In addition, the results of a large-scale field test and a validity study were presented. A major strength of the CD-CAT system is that the items were constructed in advance according to predetermined attributes. Such an attribute-based item writing method makes it possible to produce a sufficient number of items for each attribute being measured. The validity study indicates that the external measure of students' English academic performance is highly consistent with the CD-CAT's model-estimated skills mastery profiles. According to cognitive interviews of a group of three teachers, the diagnostic feedbacks generated by CD-CAT were considered useful for future remedial purposes. The

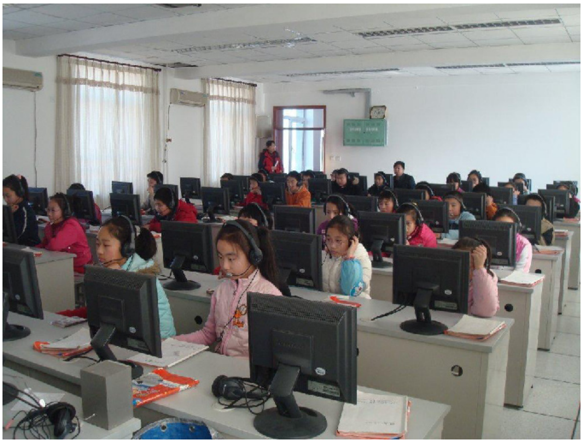


Figure 3. In January 2011, about 30,000 Grade 5 Students in Dalian, China, took a CD-CAT for the English Proficiency Level II Assessment. The Web-delivered testing lasted three days, with a peak of 2,000 students taking the test simultaneously. In the picture, the students are taking CD-CAT by using their school's PC's connected to the Internet.

teachers either agreed or strongly agreed that the cognitive feedback provides valuable information about students' strengths and weaknesses.

The current study showed that the DINA model effectively identified the sequential latent class of each student and assessed the extent to which the items were informative of the attributes. Therefore, the potential significance of the application lies in the evidence that both achievement levels and skill-mastery levels can be sequentially estimated in a CD-CAT setting.

Even though the DINA model worked well in the current study, many other models should be included in the future studies. The nature of how cognitive attributes interact with each other to arrive at a response to an item should be studied. Toward this end, several other models may generate better results, such as the Fusion model (Hartz, Roussos, and Stout 2002), NIDA model (Maris 1999), Hierarchical DINA model (de la Torre and Douglass 2004), DINO model (Templin and Henson 2006), Multicomponent Latent Trait model (Embretson 1985), and others.

In China, a big challenge to bring CAT to schools is the affordability of hardware, software, and professional testing sites. To this end, the current study sets a good example that a large scale CD-CAT implementation can be based on the cutting-edge Browser/Server Architecture that is indeed a cost-effective and user-friendly alternative to the more-traditional Client/Server design. The B/S architecture does not require specialized client software, extensive additional hardware, or detailed knowledge of the network environment.

Recently, a large scale web-based online CAT was carried out in Dalian, China, in January, 2011. About 30,000 students participated in this three-day continuous assessment with a maximum of 2,000 students taking the test simultaneously. About 2,000 PCs owned by the local schools were connected via the Internet. Figure 3 shows a group of students taking the CD-CAT test using their school's PC network connection. The PCs successfully served as test-taking terminals and B/S architecture enabled the central processing units (CPUs) to stick together so as to form a gigantic computing force to make such large scale testing flawless. Given the scope of such large scale CD-CAT application, more research will be needed in the near future.

Appendix 1

Attributes of Defining the English Level-2

Index	Name of attribute	Specification
A1	Reorganization of words	Students can recognize words and phases.
A2	Understanding of words	Students can understand meanings of words and phases and can use in their context.
A3	Understanding of grammar	Students can recognize grammar knowledge in the relative context, and can correctly judge and select.
A4	Obtaining direct information after listening	Students can understand sentence they listened; Students can understand simple dialogs they listened by supporting with short words, accurately capture particular information directly given by the dialogs.
A5	Responding after listening to the communication language	Students can understand the communication language they listened and response accurately.
A6	Obtaining indirect information after listening	Students can listen to dialogs and discourses and understand the content listened by simply judgment and inference etc.

A7	Obtaining direct information by reading	Students can understand simple stories and short passages they read, and find out particular information directly described in the stories and short passages.
A8	Obtaining indirect information by reading	Students can understand simple stories and short passages they read, and analyze the information which are not directly given in the short passages and stories by judgment and inference etc.

References

BOCK, D., and MISLEVY, R. (1982), "EAP Estimation of Ability in a Microcomputer Environment", *Applied Psychological Measurement*, *6*, 431–444.

CHENG, Y., and CHANG, H. (2007), "The Modified Maximum Global Discrimination Index Method for Cognitive Diagnostic Computerized Adaptive Testing[R]", in *Proceedings of the 2007 GMAC Computerized Adaptive Testing Conference*, June 7, ed. D. Weiss.

CHENG, Y. (2009), "When Cognitive Diagnosis Meets Computerized Adaptive Testing: CD-CAT", *Psychometrika*, *74*, 619–632.

CUI Y., LEIGHTON J.P, and ZHENG Y.G. (2006), "Simulation Studies for Evaluating the Performance of the Two Classification Methods in the AHM", paper presented at the Annual Meeting of National Council on Measurement in Education, San Francisco, CA.

DE LA TORRE, J., and DOUGLAS, J.A. (2004), "Higher-Order Latent Trait Models for Cognitive Diagnosis", *Psychometrika*, *69,* 333–353.

DE LA TORRE, J., and DOUGLAS, J.A. (2008), "Model Evaluation and Multiple Strategies in Cognitive Diagnosis: An Analysis of Fraction Subtraction Data", *Psychometrika , 73,* 595–624.

DE LA TORRE, J. (2009), "DINA Model and Parameter Estimation: A Didactic[J]", *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

DE LA TORRE, J., and LEE, Y.S. (2010), "A Note on the Invariance of the DINA Model Parameters", *Journal of Educational Measurement*, *47*, 115–127.

DOEY, L., and KURTA, J. (2011), "Correspondence Analysis Applied to Psychological Research", *Tutorials in Quantitative Methods for Psychology*, *7(1),* 5–14.

EDUCATIONAL TESTING SERVICE (2004), *Arpeggio: Release 1.1* [Computer Software], Princeton, NJ: Author.

EMBRETSON, S. (1995), "A Cognitive Design System Approach to Generating Valid Tests: Application to Abstract Reasoning", *Psychological Methods, 3*, 300–396.

HARTZ, S. (2002), "A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality", unpublished doctoral thesis, University of Illinois at Urbana-Champaign.

HARTZ, S., ROUSSOS, L., and STOUT, W. (2002), *Skill Diagnosis: Theory and Practice* [Computer software user manual for Arpeggio software], Princeton, NJ: Educational Testing Service.

JANG, E. (2008), "A Framework for Cognitive Diagnostic Assessment", in *Towards an*

Adaptive CALL: Natural Language, eds. C.A. Chapelle, Y.-R. Chung, and J. Xu, pp. 117–132.

JUNKER, B., and SIJTSMA, K. (2001), "Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory", *Applied Psychological Measurement*, *25(3):* 258–272.

LEIGHTON, J.P., and GIERL, M.J. (eds) (2007), *"Cognitive Diagnostic Assessment for Education: Theory and Practices"*, Cambridge: Cambridge University Press.

LEIGHTON, J.P., GIERL, M.J., and HUNKA, S.M. (2004), "The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach", *Journal of Education Measurement*, *41(3)*: 205–237.

MACREADY, G.B., and DAYTON, C.M. (1977), "The Use of Probabilistic Models in the Assessment of Mastery", *Journal of Educational Statistics, 2,* 99–120.

MARIS, E. (1999), "Estimating Multiple Classification Latent Class Models", *Psychometrika*, *64*, 187–212

MCGLOHEN, M., and CHANG, H. (2008), "Combining Computer Adaptive Testing Technology with Cognitively Diagnostic Assessment", *Behavior Research Methods 40(3),* 808–821.

MCGLOHEN, M. (2004), "The Application of Cognitive Diagnosis and Computerized Adaptive Testing to a Large-Scale Assessment", unpublished doctoral thesis, University of Texas, Austin, TX.

MISLEVY, R. (1994), "Probability-Based Inference in Cognitive Diagnosis", Education Testing Service Research Report, Princeton, NJ.

TATSUOKA, C. (2002), "Data Analytic Methods for Latent Partially Ordered Classification Models", *Journal of the Royal Statistical Society*: *Series C (Applied Statistics), 51(3),* 337–350.

TATSUOKA, C., and FERGUSON, T. (2003), "Sequential Analysis on Partially Ordered Sets", *Journal of the Royal. Statistical Society Series B, 65*, 143–157.

TATSUOKA, K.K. (1983), "Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory [J]", *Journal of Educational Measurement, 20(4),* 345–354.

TATSUOKA, K.K. (1995), "Architecture of Knowledge Structures and Cognitive Diagnosis: A Statistical Pattern Recognition and Classification Approach*",* in *Cognitively Diagnostic Assessment,* eds. P.D. Nichols, S.F. Chipman, and R.L. Brennan, Hillsdale: Lawrence Erlbaum Associates, pp. 327–361.

TATSUOKA, K.K., and TATSUOKA, M.M. (1997), "Computerized Cognitive Diagnostic Adaptive Testing: Effect on Remedial Instruction as Empirical Validation", *Journal of Educational Measurement*, *34(1),* 3–20.

TEMPLIN, J. (2005), *Arpeggio 2.0*, [Computer software and Manual], Author owned: retrievable upon request.

TEMPLIN, J., HENSON, R., and DOUGLAS, J. (2006), "General Theory and Estimation of Cognitive Diagnosis Models: Using Mplus to Derive Model Estimates", paper presented at the April 2007 National Council on Measurement in Education training session, Chicago, IL.

VON DAVIER, M. (2005), "General Diagnostic Model Applied to Language Testing Data", Education Testing Service Research Report, RR-05-16, Princeton, NJ.

WEISS, D. (1983), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing,* New York: Academic Press.

XU, X., and VON DAVIER, M. (2006). "Cognitive Diagnosis for NAEP Proficiency Data", Education Testing Service Research Report, Princeton, NJ.

XU, X., CHANG, H., and DOUGLAS, J. (2005), "Computerized Adaptive Testing Strategies for Cognitive Diagnosis", paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

XU, X., CHANG, H., and DOUGLAS, J. (2003), "A Simulation Study to Compare CAT Strategies for Cognitive Diagnosis", paper presented at the March annual meeting of National Council on Measurement in Education, Chicago, IL.

XU, X., and DOUGLAS, J. (2006), "Computerized Adaptive Testing Under Nonparametric IRT Models", *Psychometrika*, *71*, 121–137.

YANG, X., and EMBRETSON, S. (2007), "Construct Validity and Cognitive Diagnostic Assessment", in *Cognitive Diagnostic Assessment for Education: Theory and Applications[C]*, eds. J. Leighton, and M.J. Jierl, Cambridge: Cambridge University Press, pp. 119–145.

ZHU, Y., DING, S., ZHAO, T., and XU, Z. (2008), "A Polytomous Extension of AHM and a New Classification Method", paper presented at the Annual Meeting of National Council on Measurement in Education, New York, NY