Towards Personalized English Learning Diagnosis: Cognitive Diagnostic Modelling for EFL Listening

Xiaomei Ma & Yaru Meng*
School of Foreign Studies, Xi'an Jiaotong University, Xi'an, China 710049
**Email : maryann {at} mail.xjtu.edu.cn*

ABSTRACT: ***This paper aims to discuss how an EFL listening diagnostic model is constructed based on the theory of Cognitive Diagnostic Assessment (CDA). Compared with the traditional classical assessment, CDA can offer fine-grained feedbacks on the learner's knowledge structure and cognitive process. The diagnostic test model construction starts with the identification and definition of EFL attributes relevant to listening comprehension, followed by the hypothetic Q-matrix model and diagnostic test design, and finally, a psychometric G-DINA model analysis. The validated diagnostic model can offer reliable diagnostic feedbacks for group as well as individual levels both for listening competence and attribute mastery profiles, and is therefore qualified for online diagnostic purposes.***

Keywords: Personalized learning, EFL listening comprehension, Cognitive diagnostic assessment, Diagnostic model

1. INTRODUCTION

Personalized EFL learning is a learner-cantered modern education approach emphasizing "the tailored EFL instruction". It affords the learner a degree of choice with respect to what is learned, when it is learned and how it is learned (Hargreaves, 2009). It provides learners the opportunity to learn in ways that suit their individual learning styles and multiple intelligences. However, how to precisely identify learners' individual differences, and their gaps in language learning, so as to provide them with "tailored" instruction is a challenging question that language researchers and teachers are endeavouring to answer. For many years, suggestions have been made that intelligent tutoring technology should be used to model learners' knowledge structure (Chapelle, 2010) or intelligent assessment be constructed for analysis of learners' constructed responses (Alderson, 1988, Corbel,1993, cited in Chapelle,2010). However, these aims seem far from being achieved since the research becomes extremely complex, crossing the boundaries between assessment, language and technology (Chapelle, 2010).

Cognitive diagnostic assessment (CDA) is a newly developed psychometric approach, once empowered by Internet technology, makes it possible to create an intelligent response model that provides learners with "personalized" or "tailored" EFL learning anytime and anywhere. Unlike traditional assessment, CDA aims to measure fine-grained information of individual learners' knowledge states and cognitive processing skills well beyond an overall score. With CDA, learners' latent knowledge structures and cognitive learning skills could be diagnosed through their item response patterns, and thus the individuals' strengths and weaknesses could be assessed, and more importantly, it is followed by detailed diagnostic feedback from different perspectives (Leighton & Gierl, 2007 , Rupp, Templin & Henson 2010).

Compared with other skills, EFL listening is the most difficult skill (Graham, 2006) in some sense, and merits more analysis and support (Vandergrift, 1999; Liao, 2009). Assessment experts like Buck (2011) as well as Alderson (2005) articulate the acute need for the creation of new diagnostic listening assessments that will identify specific areas where learners need improvement, and in so doing will better inform the instructional process regarding learners' listening abilities. This study integrates the theories of CDA with cognitive theories of EFL listening to construct a diagnostic test model (DTM). The constructed DTM is then employed in the web system as an assessing instrument for the learners to diagnose their deficiencies in listening and for teachers to work out remedial instructions as well. Specifically, this paper aims to present how the listening diagnostic test model is designed and verified using cognitive diagnostic approaches.

2. LITERATURE REVIEW

2.1 Theoretical Basis

2.1.1 Cognitive theories in listening comprehension

Listening comprehension is a highly complex process of information encoding and decoding. Though it may seem effortless for native listeners, successful comprehension in foreign language is actually the result of a myriad of complex cognitive processes because of its transient nature and the limited degree of control by the listener on the stream of speech. According to Buck (2011), listening assessments should be designed so that they can be used diagnostically to evaluate and monitor learners on particular aspects of their language skills. Measuring how the process

works involves interdisciplinary knowledge in linguistics, cognitive psychology and psychometrics. Buck (2011) points out that listening requires both linguistic and non-linguistic knowledge. Among linguistic knowledge are phonology, lexis, syntax, semantics and discourse structure and non-linguistic knowledge includes topics, the context and general knowledge about the world and how it works. Apart from these, complex mental processes are also highly involved.

From cognitive perspective of Three-phase Language Comprehension Model (Anderson's 2000), language comprehension goes through three levels of processing: perception, parsing and utilization. The model provides information on how aural texts are processed and comprehended in human brains. The perceptual process is the first stage by which the acoustic message is originally encoded. In listening, this process involves segmenting phonemes from the continuous speech stream. In the parsing stage, the segmented words are transferred into a mental representation of the combined meaning of the words. In the utilization stage, listeners relate a mental representation of the text to existing knowledge which is stored in long-term memory to get a meaningful understanding of the whole message. The three phases are interrelated and recursive in language comprehension. Constructing diagnostic test model must take into account the complex cognitive processes underlying listening comprehension.

2.1.2 Cognitive Diagnostic Assessment (CDA)

As a newly developed psychometric theory, Cognitive Diagnostic Assessment is developed from the combination of cognitive psychology and measurement theory. It aims to measure the specific knowledge states and cognitive processing skills a test taker has acquired (Leighton & Gierl, 2007). Unlike a traditional test, which merely provides an overall score without indicating in which specific areas the test takers are weak, CDA test can specify students' latent proficiency, or potential knowledge structure underlying the overall test score. This specification allows for possible intervention to address individual and group needs and improve instruction for students' effective learning and progress (Lee, 2009). CDA models generally follow four major steps: (a) definition of attributes, (b) Q-matrix construction, (c) data analysis and validation, and (d) score reporting and feedback. In order to understand CDA, basic concepts such as Attributes, Q-matrix, Attributes Master Pattern should be clarified.

Attributes

In CDA, attributes refer to cognitive processes, strategies, skills, and any knowledge components of the test item (Birenbaum &Tatsuoka, 2005; Lee & Sawaki, 2009, Rupp, Templin & Henson, 2010). For example, sound discrimination, word recognition or gist understanding can be taken as cognitive or linguistic attributes in listening comprehension. On the other hand, selective attention, short term memory and note taking can serve as cognitive or strategic attributes. However, the number of attributes should be kept within a manageable degree, say, no more than 15 for the language diagnosis (Lee & Sawaki, 2009). Too many fine grained attributes would lead to rather complex statistical processing.

Q-Matrix

A Q-matrix is a two-dimensional incidence matrix used to reveal the relationship between attributes and a particular set of test items (Rupp, Templin & Henson, 2010). To be specific, it is about whether mastery of an attribute is required by an item. A Q-matrix then is constructed with items in the row and attributes in the column, as Table 1 illustrates. Its entries can be expressed with 0 or 1 indicating whether or not a particular attribute is involved in the cognitive process when students respond to an item. For instance, the Q-matrix below (Table 1) exemplifies a listening test with five items (in five rows) and three attributes (in three columns). Item 1 involves only attribute one (A1 sound discrimination) whereas Item 2 is supposed to test attribute two and three (A2 word recognition and A3 main idea understanding) and item 5 all the three attributes (A1, A2 and A3). So, in order to answer item 5 correctly, the student has to master all three attributes.

On the surface, a Q-matrix embodies the design of the assessment instrument in use and in essence determines the quality of the diagnostic information obtained through the assessment instrument (Rupp & Templin, 2008). It goes beyond the item scores, and probes into the underlying knowledge structure and cognitive process of both the test items and the respondents. So it is crucial to evaluate whether the Q-matrix is reasonably constructed. Otherwise an unreasonable Q-matrix may lead to a wrong diagnosis (Zhang, 2006; Tu, Qi & Dai, 2008).

Table 1 Example of a Q-matrix

	A1 (Sound discrimination)	A2 (Word recognition)	A3 (Main idea understanding)
Item 1	1	0	0
Item 2	0	1	1
Item 3	0	0	1
Item 4	0	1	0
Item 5	1	1	1

Attribute Mastery Pattern

Attribute mastery patterns, also referred to as knowledge state, consist of various combinations of all attributes involved in the test items. CDA model analysis of the test takers' responses will classify them into different mastery groups. Each group represents a different set of attributes. Table 2 illustrates the expected attribute mastery patterns of a test with 3 attributes. Test takers falling into Pattern 1 have not mastered any attributes (3 "0"s) while test-takers with Pattern 4 mastered only A3 (Attribute 3). In contrast, test takers with Pattern 8 have mastered all the three attributes (3 "1"s).

Table 2 Example of Attribute Mastery Patterns

Pattern Number	Representation of Pattern
Pattern 1	0,0,0
Pattern 2	1,0,0
Pattern 3	0,1,0
Pattern 4	0,0,1
Pattern 5	1,1,0
Pattern 6	1,0,1
Pattern 7	0,1,1
Pattern 8	1,1,1

This classification of test takers into latent groups based on attribute mastery patterns contributes immensely to better understanding of learners benefiting both learning and teaching.

2.2 Previous Studies on CDA

Studies on CDA arose in late 20th century and have been growing rapidly in recent years in pyschometrics. Most of them focus on theoretical issues of psychometric models and model justification (Tatsuoka, 1983; de la Torre, 2008a; Rupp et el, 2010, De Carlo, 2011). To date, over 120 models have been developed which demonstrates the fast progress in CDA theory (Fu & Li, 2007, cited in Lee & Sawaki, 2009). Among these, G-DINA model, the general version of DINA (Deterministic Input, Noisy and Gate) has been widely accepted for its simplicity of computation and estimation in identifying the role an individual attribute plays in completing a task. The concept of Q-matrix (Tatsuoka, 1983) has also driven the CDA theory into a new phase, including validation of Q-matrix specification (Rupp & Templin, 2008; DeCarlo, 2011), analysis of model goodness of fit (Sinharay & Almond, 2007; de la Torre, 2011), and estimation algorithm (de la Torre, 2008a). However, the theoretical research on CDA models seems far removed from the reach of language researchers.

CDA based language assessment is still at its infancy and it is expected to become an emerging research orientation. Kasai (1997) applied one major CDA model to the reading comprehension section of the TOEFL, identifying sixteen primary attributes and another eleven interaction attributes based on the primary attributes. Buck et al. (1997) also analyzed the cognitive attributes of a multiple-choice test of L2 reading comprehension from TOFEL based on CDA and Buck and Tatsuoka (1998) examined attributes of a free response listening test. In the past decade, various CDA models have been used in language assessment. Lee and Sawaki (2009) applied three cognitive diagnosis models to ESL reading and listening assessment. Wang, Pearson, and Gierl (2011) used the AHM (Attribute Hierarchical model, another CDA model) to make diagnostic inferences about examinee's cognitive skills in critical reading. Jang (2005, 2009a, 2009b) did studies on reading in the context of NG TOEFL (Next Generation Test of English as a Foreign Language). She argued that the construction of a Q-matrix requires multiple sources of evidence supporting the representation of the construct with well-defined cognitive skills and their explicit links to item characteristics. In China, very few studies could be found in CDA-based language assessment. Cai (2010) conducted a study based on CDA to assess the group-level EFL reading problems of middle school students. Another study is the application of diagnostic reading attributes to evaluate reading abilities for hundreds of middle school students (Cai, Ding, Tu, 2011).

The previous CDA-based studies in language indicate that most of the diagnostic tests focus on L2 reading from large scale tests like TOFEL or SAT, mainly from the perspective of psychometrics. Very few attempts were made to study EFL listening cognitive diagnosis from linguistic perspective, let alone in developing Internet-based diagnostic tests.

3. RESEARCH DESIGN

The current paper, as a pilot study, makes an attempt on how a cognitive diagnostic test model in EFL listening is constructed and verified so as to produce accurate and reliable diagnosis. The specific research questions are as follows:

1) How is an EFL listening hypothetic diagnostic test model (DTM) constructed?
2) Is the constructed DTM fit with learners' item response data?
3) What diagnostic feedbacks can the DTM produce?

The study was conducted in the following 2 parts: (1) Hypothetic DTM construction, (2) Hypothetic DTM validation. As mentioned above, the Hypothetic DTM construction started with attribute identification, then Diagnostic Q-matrix Model Construction and test development; Hypothetic Model validation follows the procedure of test item analysis and Q-matrix model verification. 3 versions of Q-matrix were verified: one from test developers (M1), another from the domain experts (M2) and the third one was the synthesized one (M3) from M1 and M2, and G-DINA model analysis was carried out to each of them(see Figure. 1).

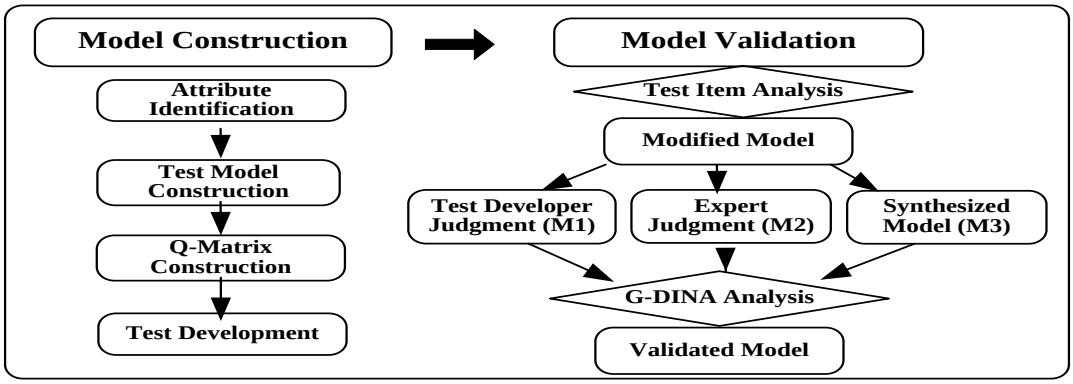

Figure 1 Procedure of Hypothetic Diagnostic Test Model Construction

3.1. Hypothetic Model Construction

3.1.1 Attribute Identification

Listening involves physiological and cognitive processes at different levels as well as attention to contextual and "socially coded acoustic clues" (Swaffar & Bacon, 1993) apart from its linguistic nature. Thus the major attributes of EFL listening comprehension can be divided into four basic categories: (1) Basic EFL linguistic knowledge including segmenting phonemes from continues speech stream, lexical and syntactical meaning; (2) Understanding content, form and function in sentence level; (3) Understanding content, form and function above sentence level; (4) Inferring the implied meaning in sentence level or above sentence level (Wang, Mark & Gierl, 2011).

Of the above 4 basic categories, more specific candidate attributes (over 30) were sorted out. 25 teachers were invited as an expert panel to rank the candidate attributes(A) from the most important to the least important. Eventually, 9 attributes were chosen as factors that affect listening comprehension. They are: A1 Phonology, A2 Lexis (Vocabulary & Phrases), A3 Syntax (Key structures and functions), A4 Details & facts, A5 Main idea, A6 Inference, A7 Background knowledge, A8 Selective attention, and A9 Short term memory and note taking (See Table 3).

In addition, based mainly on Jack Richard's (1983) framework on listening skills and Buck's (2011) construct definitions on listening test, Zou's (2005) three dimensions in listening design principles were believed to be more applicable to the Chinese learners' situation. As a consequence, a theoretical framework of EFL listening attributes was developed including four dimensions into which the 9 attributes fall. They are: 1) micro-linguistic meaning comprehension, including basic language knowledge such A1, A2, and A3; 2) direct meaning comprehension such as understanding content, form and function as surface level comprehension in which A4 and A5 are included ; 3) indirect meaning comprehension like the implied meaning belongs to deep level comprehension, such as A6 and A7 ; 4) major strategies use like selective attention, short term memory and note taking (A8 and A9), (Zou, 2005, Ma & et.al, 2012). See Table 3 below for details.

Table 3 Attributes in EFL Listening Comprehension

Major dimensions	Attribute	Definition of Attributes
Micro-Linguistic Meaning	**A 1**	**Phonology**: Understanding information through phonological knowledge and prosodic features such as discriminating phonemes of liaison and assimilation, stress and weak syllable, and sentence intonation.
	A 2	**Lexis:** Understanding word meaning of low frequency, recognizing idiomatic oral expressions and chunks and guessing unfamiliar words from context.
	A 3	**Syntax:** Understanding grammatical key structures and functions, esp. subjunctive mood, subordination, emphatic construction, recognizing numbers and relevant calculation.
Direct Meaning	**A 4**	**Details & facts**: Understanding details such as time, place and the relationship between speakers.
	A 5	**Main idea**: Understanding the author's purposes, goals and strategies
Indirect Meaning	**A 6**	**Inference**: inferring implied meaning or guessing from context;
	A 7	**Background knowledge**: Understanding implied meanings through activating background, esp. cultural knowledge;
Major Strategies	**A 8**	**Selective Attention**: Knowing how and when to give selective attention.
	A 9	**Short term memory & note taking:** Knowing how to memorize key information through effective note-taking.

3.1.2 Diagnostic Test Model (Q-Matrix)

With the above framework and diagnostic attributes as a guideline, the details of the test model were established. It reflects the nature of listening comprehension, listening strategies and skills as well as the requirements of college English syllabus (Band 4).

Table 4 The Blueprint for DTM

Section	Item No.	Attributes & Testing tasks for different purposes	Task types	Cognitive phase
I	1.	(A1)Phonology: discriminating phonemes of liaison and assimilation	Statement with MCQ	Perception
	2.	(A1) Prosody: stress and weak form, intonations		
	3.	(A2) Recognition of words	Short conversation with MCQ	Perception + Parsing
	4.	(A2) Recognition of idiomatic oral expressions and chunks		
	5.	(A2) Guessing of new words from the context		
	6.	(A3) Grammatical functions, Subjunctive mood and inverted		
	7.	(A3) Recognition and Calculation of numbers		
	8.	(A5) Understanding of main idea, Comprehension of main idea of short conversation		Parsing + Utilization
	9.	(A6) Making inference		
	10.	(A7) Making inferences based on cultural knowledge		
II	11.	(A7) Making inferences based on background knowledge	Short passage with MCQ	Utilization
	12.	(A5) Comprehension of main idea of short passages		
	13.	(A4) Comprehension of details and facts in short passages		
	14.	(A6) Making inference in short passages		
III	15.	(A9) Note-taking, working memory, choosing the exact missing sentence	Dictation with MCQ	attention
	16.	(A8) Note-taking, working memory, choosing main idea of the miss sentence		memory
IV	17.	(A5)Understanding of main idea of a video clip	Video clips with MCQ	Utilization
	18.	(A4)Facts and details in video clips		
V	19.	Self-report question of difficulty	Self-report MCQ	Emotion
	20.	Self-report question of affect & anxiety		

Note: A=attribute A1=attribute 1 MCQ=multiple choice question

The test model becomes the essence of the blueprint according to which, the test items were developed (see Table 4). This blueprint not only demonstrates the test Q-matrix by specifying the relationship between attributes and test items, but also the relationship between task types and phases in cognition process. It provides useful information for the test writers to develop tests.

From the above test blueprint derives the Q-Matrix model as an ideal response model (See Table 5) whether it is reasonably designed needs further verification from other evidence like domain experts.

Table 5 Hypothetic DTM Q-matrix M1 (Test Developers' Model)

Item	A1	A2.	A3	A4	A5	A6	A7	A8	A9
1	1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0
6	0	0	1	0	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0	1	0	0	0
10	0	0	0	0	0	0	1	0	0
11	0	0	0	0	0	0	1	0	0
12	0	0	0	0	1	0	0	0	0
13	0	0	0	1	0	0	0	0	0
14	0	0	0	0	0	1	0	0	0
15	0	0	0	0	0	0	0	0	1
16	0	0	0	0	0	0	0	1	0
17	0	0	0	0	1	0	0	0	0
18	0	0	0	1	0	0	0	0	0

Notes: A= attribute required for answering each item; 1=Presence of a certain attribute; 0= Absence of a certain one

3.1.3 Test Item development

Based on the above attributes and Q-Matrix, 4 test developers began to write test items. As indicated above, the model test is composed of 5 parts, with 18 items in the first 4 parts and 2 self-report ones in the last complementary part. All the test items take the form of multiple choice questions. Except for the two short video clips (item 17 and 18), all the listening materials were recorded by two native speakers, one male from America and the other female from Britain. Vocabularies are controlled at the level of College English Test Band 4 according to the glossary of College English Curriculum Requirements (2007). The test types include short conversation, short passage, long passage, dictation and video clip, of genres typical of CET tests such as lectures, campus life, and so on (See Table 4). Eventually, 180 items in 10 tests were produced strictly following the definition of 9 attributes and the test blueprint.

3.2 Hypothetic DTM (Q-Matrix) Validation

In order to verify whether the test model was good enough to serve as a diagnostic test instrument, the analysis of item parameters (based on item response theory) was made first and then followed by the G-DINA model-data fit analysis.

3.2.1 Data Collection

The data were learner's pilot responses of the first test. The piloting was administrated in attempt to check the involved assessment dimensions and to uncover the diagnostic information embedded in examinees' item response data so that inferences about their strengths and weakness can be made. The piloting sample included 87 paper-based participants and 254 online-based participants, totally 341. They were all first year college students in a Northwest China "985" university.

3.2.2 Item Analysis

Item analysis aims to identify poorly written items and improve the quality of items on a test. The data were processed using BILOG-MG. Interpretations of the results were made to screen out the unqualified items and choose the ones with the best statistical indexes. In item analysis, item parameters such as discrimination(a), difficulty(b) and guessing probability(c) can be obtained (but only the first two parameters can be obtained at the first attempt). It is noted that the mean discrimination (a=0.40) and difficulty (b=-0.85) all fall into the acceptable value ranges 0 to 3 and -3 to 3

respectively (Luo, 2012) and discrimination is often better between 0.30 and 2. However, statistical analysis showed that discrimination of item 2 is exceptionally low (a=0.25) and the difficulty of item 15 is over 3 (b=4.08). The two items (shaded in table 6) interfere with the reliability of the test as a whole (See Table 6).

Table 6 Item parameters of the DST Model (18 items for 2 parameters)

Item	a	b	Item	a	b
No.01	0.41	-2.21	No.10	0.46	-1.19
No.02	0.25	1.07	No.11	0.44	-0.88
No.03	0.50	-1.77	No.12	0.37	0.23
No.04	0.40	-1.47	No.13	0.35	-1.55
No.05	0.35	-1.58	No.14	0.32	-0.02
No.06	0.50	-2.14	No.15	0.18	4.08
No.07	0.37	-2.27	No.16	0.33	1.02
No.08	0.59	-0.78	No.17	0.32	-2.85
No.09	0.64	-1.42	No.18	0.35	-1.63

Note: a=discrimination b=difficulty

After item 2 and 15 were removed, the item analysis results improvement was statistically significant. Table 7 contrastively shows the mean and standard deviation between the 18 item test on the left side of Table 7 and the 16 item test on the right side.

Table 7 Contrastive parameters of the 2 DTM of 18 items and 16 items

18 Items （ 2 Parameters）			**16 Items （ 3 Parameters）**		
Parameter	Mean	S. D.	Parameter	Mean.	S. D
a	0.40	0.11	a	0.81	0.37
b	-0.85	1.64	b	0.41	0.95
c	—	—	c	0.47	0.05

Through item parameter analysis, the adjusted test model with 16 items proved to be good enough to make the further G-DINA Model analysis.

3.2.3 Q-matrix Model Verification

The original Q-matrix (M1) was built by the 4 test developers (See the left column of Table 8). Their EFL teaching experience is 17, 18, 20 and 33 and two of them have received professional training in test development. The constructed M1 was then verified by domain experts consisting of 8 senior English teachers, all of whom hold MA degrees in Applied Linguistics. Their teaching experience ranges from 13 to 21 years (see the right column of Table 8).

Table 8 Information of 4 Test Developers and 8 Domain Experts

4 Test Developers		**8 Doman Experts**			
Name	**TE (year)**	**Name**	**TE (year)**	**Name**	**TE (year)**
1 TD No. 1	33	1 DE No. 5	20	5 DE No. 9	21
2 TD No. 2	17	2 DE No. 6	19	6 DE No. 10	17
3 TD No. 3	18	3 DE No. 7	19	7 DE No. 11	13
4 TD No. 4	20	4 DE No. 8	19	8 DE No. 12	13

TD=Test Developer; DE=Doman Expert

Again, take Test 1 as a pilot study. The domain experts were required to judge the test items developed by the 4 test developers respectively and code what attributes are necessary to successfully complete each item. One attribute was determined when 4 or more experts agreed upon. It was then assigned/coded 1, otherwise it would be 0. Their coding result becomes the second Q-matrix model (M2). It was compared with the test developers' Q-matrix model (M1). Then, M1 and M2 were synthesized into Q-matrix model 3 (M3), the one decided by 12 teachers (4 test developers and 8 domain experts). See Table 9 and 10 for M2 and M3.

Table 9 Q-matrix Model 2 (by 8 experts)

Item	A1	A2.	A3	A4.	A5	A6	A7	A8	A9
1	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0
3	0	1	0	0	0	1	0	0	0
4	0	0	0	0	0	1	0	0	0
5	0	0	1	0	0	1	0	0	0
6	0	0	0	1	0	0	0	0	0
7	0	1	0	1	0	0	0	0	0
8	0	0	0	1	0	1	0	0	0
9	0	0	0	0	0	0	1	0	0
10	0	0	0	1	0	0	1	0	0
11	0	0	0	0	1	0	0	0	0
12	0	0	0	1	0	0	0	1	0
13	0	0	0	0	0	1	0	0	0
14	0	0	0	0	0	0	0	1	1
15	0	0	0	0	1	0	0	1	1
16	0	0	0	1	0	0	0	0	0

Table 10 Q-matrix Model 3 (by 12 teachers)

Item	A1	A2.	A3	A4.	A5	A6	A7	A8	A9
1	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0
3	0	1	0	0	0	1	0	0	0
4	0	1	0	0	0	1	0	0	0
5	0	0	1	0	0	1	0	0	0
6	0	0	1	1	0	0	0	0	0
7	0	1	0	1	1	0	0	0	0
8	0	0	0	1	0	1	0	0	0
9	0	0	0	0	0	0	1	0	0
10	0	0	0	1	0	0	1	0	0
11	0	0	0	0	1	0	0	0	0
12	0	0	0	1	0	0	0	1	0
13	0	0	0	0	0	1	0	0	0
14	0	0	0	0	0	0	0	1	1
15	0	0	0	0	1	0	0	1	1
16	0	0	0	1	0	0	0	0	0

Notes: The Test has 16 items. The boxed and shaded "1"s are the identified attributes for each item.

Each Q-matrix represents different relationship between attributes and the test items. Which of the 3 hypothetic models, is the most reasonable and valid to produce reliable diagnostic results will be determined through psychometric approach, G-DINA model analysis in this case, to see the model-data fit.

3.2.4 G-DINA Model Analysis

After item analysis, the original 18-item hypothetic model was rejected and the modified 16-item model was adopted. The collected data for the 16 items were used for G-DINA model analysis with the 3 hypothetic Q-matrixes (M1, M2 and M3) separately.

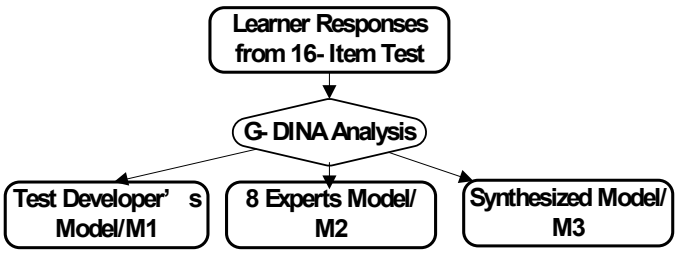

Figure2 G-DINA Analysis Procedure

G-DINA Model, one of many current CDA models, (de la Torre, 2008b) is employed in this study because of its easy recognition of the effects an individual attribute plays in completing a task. Mastery of one attribute will increase the probability of a test-takers responding correctly to an item (de la Torre, 2008b). The more attributes are mastered, the higher the possibility for testees to provide correct responses to items. For G-DINA model analysis, the software, Ox 6.21 (Doornik, 2002-2011)[1] was used to check model-data fit. The data analysis can be generally divided into two parts: one is the model-data fit statistics which includes absolute fit statistics and relative fit statistics. The other is inference of diagnostic outcomes, which involves item parameter estimates and standard errors, latent class and their posterior probabilities, and person classification (psychometric concepts). The former aims at verifying whether the results are statistically sound in diagnostic model construction i.e. Q-matrix, whereas the latter offers detailed parameters of test and learners. If the model-data fit fails, the attributes of the Q-matrix has to go through multiple rounds of adjustment and trial (for example, one attribute was decided with 4 or 5 experts/teachers' agreement) until a good fit is obtained. The purpose was to have every possible attribute considered and tried.

In the following analysis, the absolute and relative fit evaluation is integrated to judge the model-data fit. The absolute fit statistics Max.z (r) and Max.z (l) are used for determining if the Q-matrix model fits the data adequately. If

[1] Ox and all its components are owned by Jurgen A Doornik (JAD) and protected by United Kingdom and international copyright laws.

the Max Z scores of both Z (r) and Z (l) are smaller than the Zc Score (BC), the Q-matrix model is retained; otherwise it is rejected. The relative fit statistics (e.g., AIC and BIC) are only for comparing different Q-matrix models; the lesser the values the better, especially with BIC value (Chen, Torre, & Zhang,2013)(See Table 11).

Table 11 Absolute and Relative Fit Statistics

Fit statistics	**Absolute Fit Statistics**			**Relative Fit Statistics**		
	Max.z(*r*)	Max.z (*l*)		-2LL	AIC	BIC
M1	4.81	4.52		6042.76	6616.76	7716.51
M2	**3.72**	**3.53**		6033.15	7159.15	9316.50
M3	4.40	3.79		5827.77	6961.77	9134.44
	α=0.10	2.73	3.34	3.34		
Zc Score (BC):	α=0.05	2.96	3.53	3.53		
	α=0.01	3.42	**3.93**	**3.93**		

Note. M1= test developer's Matrix; M2= experts panel's Matrix; M3= the synthesized Matrix of the former two.
Max. z(*r*) = maximum z score for *r*; Max. z(*l*)= maximum z score for *l;*
-2LL=-2 log likelihood; AIC= Akaike' information criterion; BIC= Bayesian information criterion

As indicated in the above table, both maximum z score for *corr, z(r)*and maximum z score for *log, z(l)* of M2 are smaller than Zc Score (BC) at α=0.01 level, meaning the 8 domain expert- model M2 fits the data best and is accepted as the estimation and classification model. At the same time, M1 and M3 are rejected.

From the (above) G-DINA analysis, we can say that M2 has a good fit with learner response data and therefore the results of DTM M2 can be used for further diagnostic inferences. It can be concluded that the diagnostic Q-matrix model (M2) by the domain experts is proved to be reliable and valid for further online application.

4. DTM DIAGNOSTIC RESULTS

By applying G-DINA model, we can find that the statistic output presents not only the results of the model-data fit, but also very comprehensive estimation outcomes. Among all the parameters and statistic results, person classification is the attribute classification of learners both on group level and individual learner and attribute level.

4.1 On group level

Person classification offers an overall picture of the 9 attributes mastery patterns both in frequency and percentage. 341 samples are classified into 122 patterns, with 50% learners falling within the top 20 patterns (as is shown in Table 12). It can be seen that 19 test-takers who had mastered all the attributes account for the top 5.56% of the 341 students, followed by 4.97% of the learners who mastered all but attribute 9 (note taking and memory). 4.39% (15 students) fall into the third pattern, mastering 6 attributes. It is clear that with the accurate information of learners' different attribute mastery levels, the teacher will know what the students' gaps are, and what remedial measures should be taken to bridge them. The G-DINA Model analysis on person classification is evidently significant for pedagogical improvement.

Table 12 Attribute Mastery Patterns for Top 20 (50% of the sample)

Person	classification	Frequency	Percent%	Person	classification	Frequency	Percent %
(1)	1 1 1 1 1 1 1 1 1	19	5.56	(11)	1 1 1 1 0 1 1 1 0	7	2.05
(2)	1 1 1 1 1 1 1 1 0	17	4.97	(12)	1 0 1 1 1 0 1 1 0	7	2.05
(3)	1 0 1 0 0 1 1 1 1	15	4.39	(13)	1 0 1 0 0 0 1 0 0	6	1.75
(4)	1 1 1 1 0 0 0 1 0	12	3.51	(14)	0 0 1 0 1 0 1 0 0	6	1.75
(5)	1 1 1 1 1 0 1 1 0	10	2.92	(15)	1 1 1 1 0 0 1 0 0	5	1.46
(6)	1 1 1 1 0 0 1 1 0	10	2.92	(16)	0 0 1 0 0 1 1 1 1	5	1.46
(7)	1 0 1 1 0 0 1 1 0	10	2.92	(17)	1 1 1 0 1 0 0 0 0	4	1.17
(8)	1 1 1 1 1 0 0 1 0	9	2.63	(18)	1 1 0 1 1 1 1 1 0	4	1.17
(9)	1 1 1 1 0 1 1 1 1	9	2.63	(19)	1 1 0 0 0 1 0 0 1	4	1.17
(10)	1 1 1 0 0 0 1 1 0	8	2.34	(20)	1 0 1 1 1 1 1 1 1	4	1.17

Note: 1=Mastery of a certain attribute; 0= Non-mastery of a certain one

However there is still room for improvement. Although the 9 attributes are proven to work, to process 122 attribute mastery patterns will take quite a while to work out online. Therefore, 9 attributes will be modified if the model is uploaded onto the system.

4.2 On individual attribute level

Examining the mastery of knowledge states in listening comprehension, we can see learners' various mastery levels of individual attribute. It can be seen from Table 13 that Attribute 1, (sound discrimination) and Attribute 3 (functional sentence structures) are the skills best mastered, with A1 included in 18 mastery patterns, into which 46.8% of the test-takers fall, and A3 in 16 patterns covering 47.7% of the sample. This result suggests that A1 and A3 are the most basic attributes and should be mastered before other attributes. It is interesting to find that Attribute 8 (selective attention) and Attribute 7 (cultural background) are the 3rd and 4th better mastered skills with 42.7% and 41.5% of the sample respectively, meaning these students are better prepared in test-taking techniques, and their target cultural knowledge is fairly good for listening comprehension purpose. Attribute 4 (understanding of facts and details) and Attribute 2 (less frequent vocabulary and oral expressions) are both in 13 mastery patterns with respectively 36% and 34.5% of the sample. Attribute 9 (short term memory and note-taking skill) in 9 mastery patterns with 33.6% of the sample needs some instruction and practice. In sharp contrast, Attribute 5 (main ideas) with 23.5% of the sample, and attribute 6 (inferring implied meaning) with 24.6% of the sample, are the least mastered skills (See Table 13)

Table 13 Mastery State of Each Attribute

Attribute	A1	A2	A3	A4	A5	A6	A7	A8	A9
Mastery Order	1	6	2	5	9	8	4	3	7
Mastery sample (%)	46.8	34.5	47.7	36.0	23.5	24.6	41.5	42.7	33.6
Mastery Pattern No.	18	13	16	13	9	9	16	15	14

We note that A1(phonology) and A3(sentence structures) are surface level skills and constitute basic linguistic knowledge based on which other skills can be acquired. However, the last two attributes, A5 (main idea) and A6 (inference), are deeper level or contextualized skills not to be mastered easily. With this information, teaching and learning can be tailored to specific needs. This is the typical characteristic feature of CDA results.

4.3 On individual learner level

With CDA analysis, each learner is offered an individual knowledge state report aside from the traditional score. Students can have a clear idea of their strengths and weaknesses. Even with the same score, the learners' mastery patterns can be totally different. Table 14 shows, for instance, that the 67 students in Example 1 had the same score, i.e., 62.5, but they fell into 37 different attribute mastery patterns. In Example 3, 39 students with the score of 81.3 fell into 23 different patterns. The implication is that it is quite possible for a number of students to have the same score but to have mastery patterns that are totally different. This may seem strange on surface. However, since each attribute is tested in no less than 3 items, it is reasonable if those who got the majority of items right but these items happened to test only one or two attributes. In this case, the test scores might be high but the attribute mastery pattern can be pretty poor, just as ID 243 illustrates. Vice versa is also true. Take ID 115 for instance, this learner may score low but got most items testing 7 attributes (the "1"s) right. In contrast, student with ID 79 may score low and got most items testing the rest 6 attributes (the "0"s) wrong.

Table 14 Examples of Different Mastery Patterns with the Same Scores

Example	Test Score	No. of Respondents	No. of patterns	Sample mastery patterns
1	62.5	67	37	ID 79: 1 0 1 1 0 0 0 0 0 ID 87: 1 1 1 1 0 0 0 1 0 ID 100: 0 1 1 0 1 0 0 0 1 ID 115: 1 1 1 0 1 1 0 1 1
2	75	45	26	ID 187: 0 0 1 0 1 0 1 0 0 ID 190: 1 1 1 1 0 0 1 1 0 ID 215: 1 1 1 1 1 0 0 1 1 ID 203: 1 0 1 0 0 1 1 1 1
3	81.3	39	23	ID 243: 0 1 1 1 0 0 0 1 1 ID 276: 1 0 0 0 1 1 1 0 1 ID 277: 1 1 1 1 1 1 1 1 0 ID 305: 1 1 1 1 1 1 1 0 1

Note: ID= a student ID number in the current sample

It is obvious from the above that the CDA-based diagnostic model can go beyond the overall scores and offer fine

grain-sized estimates and classifications of the learners, so that detailed inference about their knowledge states and skill mastery profiles can be obtained. This individualized and fine-grain sized diagnosis fulfils the purpose of personalized assessment of listening comprehension that the traditional tests can never achieve. They can be used either by the instructors for future tailored teaching or by the learners themselves to embark on their personalized and autonomous learning journey.

5. CONCLUSION

This paper presents a pilot study in constructing the listening diagnostic model for PELDiaG system using Cognitive Diagnostic Assessment . The answers to the three research questions are as follows:

1) A listening hypothetic DTM is constructed in 3 steps. It begins with the identification of 9 attributes in listening comprehension. Then the diagnostic Q-matrix model was built, the pivotal step in constructing the diagnostic test model. On the basis of this model, the test items are developed.
2) The hypothetic model is validated through item analysis, Q-matrix verification and G-DINA analysis. In the item analysis, Item 2 and 15 proved to be unqualified in the parameter of difficulty and discrimination and therefore removed from the further G-DINA analysis. For Q-matrix verification, comparative psychometric analysis of test developers' Q-matrix (M1) and domain experts' judgment Q-matrix model (M2) and the synthesized Q-Matrix (M3) was conducted by using G-DINA. Hypothetic Q-matrix M2 turned out to be the best model-data fit. It can serve as the preliminary online diagnostic assessment instrument.
3) The validated diagnostic model produces different diagnostic results apart from an overall score. They include classification of learners into groups according to their attribute mastery patterns, individual learner' knowledge state and individual and group attribute mastery profiles. These detailed feedbacks make it possible for learners to be informed of their strengths and weaknesses in listening comprehension so that they know clearly what to focus on in the future, while the teacher will know what to do in teaching based on learners' attribute mastery both on individual and group levels . With the aid of the CDA models, aspects of tailored instruction and learning may be achieved.

The constructed CDA test model is effective in providing formative diagnostic feedback through a fine-grained reporting to individual learners which the traditional assessment can never do. The availability of such diagnostic feedbacks also allows the teacher to identify the learner's specific deficiencies and to plan instruction to the needs of a particular learner. This may also facilitate learners on their journey to autonomy (Meng, 2013).

This study is significant because it is among the first very few to introduce CDA into foreign language teaching and learning. Integrated with internet technology, the produced CDA-based diagnostic model will make it possible to fulfil the aim of intelligent assessment and tutoring. This, to a certain extent, signifies a great progress toward a more personalized, self-directed and easy-accessible EFL learning and assessment. The significance also lies in the fact that it not only enriches the theory of language assessment, but also demonstrates great potentials for large-scale applications in future personalized English e-learning and e-assessment. What's more, it broadens language assessment research.

Without exception, limitations in the current study cannot be ignored. First and foremost, in the Q-matrix model construction, we didn't consider test-takers' opinions on what kind of attributes they would think is needed in the test-taking process though we have the domain experts' judgment. It is more desirable to have test-takers' verbal reports in Q-matrix construction for future reserach. As Jang (2005) argued that the construction of a Q-matrix requires multiple sources of evidence supporting the representation of the construct with well-defined cognitive skills and their explicit links to item characteristics. Second, the number of test items in this pilot test was not big enough to represent all the 9 attributes, each of which is supposed to be in no less than three items according to CDA theory. In view of CDA, the more attributes are involved, the more calculating time it may take to produce attribute mastery patterns online. Therefore, the 9 attributes will be reduced or more test items should be generated if the model is uploaded onto the system . We believe these limitations will be addressed in the future research when the formal CDA EFL listening diagnostic model is constructed. We hope our efforts in this cross-disciplinary area can inspire other in-depth exploration of the CDA-based online language assessment from researchers and teaching practitioners in L2 research field.

6. ACKNOWLEDGMENT

The research project is Funded by China National Social Science Fund (Project No.12YJA740057), China Ministry of Education of Humanities and Social Science (Project No.12BYY055) and National Project "985 Phase III".

7. REFERENCES

[1] Alderson, J. C. (2005) *Diagnosing Foreign Language Proficiency---The Interface between Learning and Assessment* [M]. Britain: Continuum. 70-78.
[2] Anderson JR. (2000). *Cognitive Psychology and its Implications* [M]. New York: Worth Publishers.

[3] Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. Language Learning, 47(3), 423-466.
[4] Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. Language Testing, 15(2), 119-157.
[5] Birenbaum, M., Nasser, F., &Tatsuoka, C. (2005). Large-scale diagnostic assessment: Mathematics performance in two educational systems. Educational Research and Evaluation, 11(5), 487-507
[6] Buck, G (2011). *Assessing Listening* [M], Beijing: Foreign Language Teaching an Research Press. 56-57
[7] Cai, Y. (2010). Group-level Ability Assessment and Cognitive Diagnosis on English Reading Problem Solving. Unpublished doctorate dissertation, JiangXi Normal University.
[8] Cai, Y. Tu, D.B. & Ding., S. L. (2010). Theory andMethod on Compilation of Cognitive Diagnosis Test. Examination Research(3): 79-92.
[9] Chapelle, C.A. & Douglas (2010), *Assessing Language through Computer Technology* [M], Beijing: Foreign Language Teaching and Research Press, 114-117.
[10] Chen, J., Torre, J. d. l., & Zhang, Z. (2013). Relative and Absolute Fit Evaluation in Cognitive Diagnosis Modeling. *Journal of Educational Measurement* 50 (2): 123–140
[11] DeCarlo LT.(2011) On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix [J]. Applied Psychological Measurement. 35(1): 8-26.
[12] De la Torre J. (2008a) An empirically based method of Q-matrix validation for the DINA model: Development and applications [J]. Journal of Educational Measurement. 45(4): 343-362.
[13] De la Torre J.(2008b) The generalized DINA model. The International Meeting of the Psychometric Society. Durham, NH
[14] De La Torre, J. (2011). The generalized DINA model framework. Psychometrika, 76(2), 179-199.
[15] De la Torre J. (2010) Estimation Code for the G-DINA Model [C]. in Rupp AA. Software for calibrating diagnostic classification models: An overview of the current state of the art. Maryland: University of Maryland.18-22.
[16] Graham, S. (2006). Listening comprehension: The learners' perspective. System, 34(2), 165-182.
[17] Hargreaves, A. & Shirley, D. (2009). *The Fourth Way: The Inspiring Future for Educational Change*. Corwin: London.
[18] Jang, E. E. (2005). A validity narrative effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL: University of Illinois at Urbana-Champaign.
[19] Jang, E. E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to languedge assessment. Language Testing, 26(1), 031-073.
[20] Jang Eunice Eunhee. (2009b) Demystifying a Q-Matrix for Making Diagnostic Inferences About L2 Reading Skills. *Language Assessment Quarterly*, 6: 210-238.
[21] Kasai M. (1997). Application of the rule space model to the reading comprehension section of the test of English as a foreign language [D]. Urbana: University of Illinois at Urbana-Champaign..
[22] Lee, Y.W. & Y. Sawaki.(2009). Cognitive Diagnosis Approaches to Language Assessment: An Overview [J]. *Language Assessment Quarterly*, , 6(3): 172-189.
[23] Leighton, J. P. & M. J. Gierl. (2007). *Cognitive Diagnostic Assessment for Education-- Theory and Application* [M]. Cambridge University Press,. 3-18.
[24] Liao, Y. F. (2009). A construct validation study of the GEPT reading and listening sections: Re-examining the models of L2 reading and listening abilities and their relations to lexico-grammatical knowledge. Teachers College, Columbia University.
[25] Luo, Z. S. (2012). Item Response Theory. Beijng: Beijing Normal University Publishing Group.
[26] Ma, X., Meng Y., He, H., & R. Liu,(2012) Personalized EFL. Audio-Visiondiagnostic model construction and system development[J] Foreign Language Education(5):59-63
[27] Meng, Y. R. (2013). Developing a Model of Cognitive Diagnostic Assessment for College EFL Listeng, Unpublished doctorate dissertation, Shanghai International Studies University.
[28] Richards JC.(1983) Listening comprehension: Approach, design, procedure. TESOL Quarterly., 17(2): 219-240.
[29] Rupp A. A & Templin J. (2008) The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model [J]. Educational and Psychological Measurement., 68(1): 78-96.
[30] Rupp, A., Templin A., J.& Henson, R.(2010) *Diagnostic Measurement :Theory, Method, and Applications* [M], New York London: the Guilford Press,. 26-48.
[31] Sinharay, S., & Almond, R. G. (2007). Assessing Fit Of Cognitive Diagnostic Models A Case Study. Educational and psychological measurement, 67(2), 239-257.
[32] Swaffar, J. &Bacon, S. (1993) Reading and Listening Comprehension: Perspectives on Research and Implications for the Classroom. In A.O. Hadley Lincolnwood, ILL: National Textbook Co., 1993. 124-55.
[33] Tatsuoka KK.(1983) Rule space: an approach for dealing with misconceptions based on item response theory [J]. Journal of Educational Measurement., 20(4): 345-354.
[34] Tu, D. , Qi, S. & Dai H..(2008)Cognitive Diagnostic Assessment in Educational Testing[J]. Testing Research. 4(4):4-15
[35] Wang, C. Mark P. & J Gierl (2011). Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Critical Reading. *Journal of Educational Measurement.* Vol.48-2:165-187.

[36] Vandergrift, L. (1999). Facilitating Second Language Listening Comprehension: Acquiring Successful Strategies. ELT Journal, 53(3), 168-176.
[37] Zhang WM.(2006) Detecting differential item functioning using the DINA model [D]. Greensboro: The University of North Carolina..
[38] Zou S. (2005) Language Assessment[M], Shanghai: Shanghai Foreign Language Education Press.