Tabaran Institute of Higher Education ISSN 2476-5880
International Journal of Language Testing
Vol. 9, No. 1, March 2019

Application of Cognitive Diagnostic Models to the Listening Section of the International English Language Testing System (IELTS)

Farshad Effatpanah[1]

Received: 12 December 2018 Accepted: 17 February 2019

Abstract

The purpose of the present study was twofold: (a) to compare the performance of six cognitive diagnostic models, including a general model (GDINA), two non-compensatory models (DINA and NC-RUM), and three compensatory models (ACDM, DINO, and C-RUM), at test level to find the best model for describing the underlying interaction among the listening attributes of the IELTS exam; and (b) to diagnose the performance of Iranian candidates in the listening section of the IELTS. To accomplish these, item responses of 310 Iranian test takers to the Listening Sub-test of the IELTS exam were analyzed. The models were first compared in terms of absolute and relative fit indices for selecting the most optimal model. The results showed that the G-DINA model was the best model with regard to all fit indices among the competing models followed by the C-RUM, ACDM, NC-RUM, DINO, and DINA. Then, the C-RUM as the best specific CDM was selected for the second phase of the study. It was found that making inference and comprehending vocabulary and syntax are the most difficult listening constituents for Iranian IELTS candidates.

Keywords: *CDMs, compensatory, non-compensatory, Q-matrix, listening, IELTS*

1. Introduction

The International English Language Testing System (IELTS) is an international standardized test which is jointly administered by the British Council, the International Development Program of Australian Universities and Colleges (IDP), now known as IDP: IELTS Australia, and the Cambridge English Language Assessment. This large-scale exam includes two versions called Academic and General Training modules and measures four language skills (e.g., reading, listening, writing, and speaking). All test takers can choose to take either the General Training or Academic module. The listening and speaking parts are the same in

[1] English Department, Mashhad Branch, Islamic Azad University, Mashhad, Iran.
Corresponding author: Email: farshadefp@yahoo.com

both modules whereas reading and writing tests are different. The General Training module is intended to measure to what extent candidates are prepared to undertake non-academic activities such as immigration and work experience in English language environments in their real life. The Academic module, on the other hand, measures the language proficiency of those test takers who desire to pursue their studies in English-speaking countries at the undergraduate or graduate level. An integral part of the exam is the Listening Sub-test which assesses test takers' comprehension of various short excerpts. There are different question types in the sub-test: multiple choice, matching, plan/map/diagram labeling, form/note/table/flow chart/summary completion, sentence completion, and short-answer questions (IELTS, 2017a). Test-takers are required to respond to the questions while the tape is playing only once. It makes the listening section intensive and demanding for examinees because they should pay simultaneous attention to three skills: listening, reading, and writing (Alavi et al, 2018). To prepare examinees for the exam and especially for the listening section, a large number of institutions offer educational programs for helping students to get through the test and administer mock/practice tests to provide test takers with appropriate and test-related feedback. However, the way they give feedback to students are not so diagnostic and detailed to allow students to inform of their strengths and weaknesses in different aspects of the cognitive domain. IELTS statistics for test taker performance in 2017 (IELTS, 2017b) show that Iranian candidates performed relatively poorly on the receptive skills (e.g., listening and reading), particularly on listening sub-test. In this regard, diagnosing Iranian test takers' listening deficiencies merits extensive investigation. By identifying problematic areas of listening, students can receive appropriate and timely feedback on their performance. Also, it enables course designers and teachers to adopt effective techniques and materials to remedy the problems students mainly face in order to develop their listening ability.

Listening comprehension (LC) is the process of extracting meaning from aural input (Snowling & Holme, 2005). Just like any other of the four language skills, LC is considered as a complex, fleeting, and multidimensional process (Britton & Graesser, 2014; Rost, 2013 Rumelhart, 1980). Researchers have explained that successful comprehension relies on a wide range of cognitive skills and linguistic knowledge, including phonology, morphology, syntax, semantics, and discourse structures (Andringa et al., 2012). A variety of listening models have been proposed to demonstrate the complicated process of listening comprehension and its relationship with a set of non/cognitive characteristics. The proposed models can be classified into two general groups (Aryadoust, 2018): general models which focus exclusively on the cognitive processes under non-assessment conditions (Buck, 2001; Chapelle, 1994, 1998; Rost, 2016; Vandergrift & Goh, 2012; Wagner, 2002, 2004; Weir, 2005); and assessment models which incorporate task-related variables and test takers' ability (Buck & Tatsuoka, 1998; Freedle & Kostine, 1996; Nissan, DeVincenzi, & Tang, 1996; Richards, 1983). The results of these studies have indicated that listening

comprehension entails a number of sub-skills which empower listeners to achieve comprehension. Many attempts have been made to describe them in terms of taxonomies of sub-skills that underlie the processes (Aitken, 1978; Carroll, 1972; Flowerdew, 1994; Hughes, 1989; Munby, 1978; Oakeshott-Taylor, 1977; Richards, 1983). Along the same lines, an emerging body of scholarships have applied different types of psychometric models to support the validity of such conjectural taxonomies (Buck, 2001; Wagner, 2004; Liao, 2007; Eom, 2008). According to Buck and Tatsuoka (1998), understanding the exact nature of what knowledge, sub/skills, and abilities are involved in second/foreign language listening comprehension would help scholars to model language processing better, build logical theories of language performance, and construct language tests which can provide diagnostic information.

More recently, cognitive diagnostic models (CDMs) have received a great deal of attention due to their capability in generating fine-grained information about the learning status of students to aid further learning and instruction (Rupp, Templin, & Henson, 2010). Unlike traditional psychometric frameworks, such as classical test theory (CTT) and item response theory (IRT), including a true score or latent trait which can be used to plot students' positions on a single proficiency continuum, CDMs provide rich diagnostic information about strengths and weaknesses of the examinee's cognitive skills (Lee, de la Torre, & Park, 2012). Multiple strategies, processes, and knowledge are assumed for students in order to respond correctly to a given test item or task (Birenbaum, Kelly, & Tatsuoka, 1993). This property enables CDMs to produce "multidimensional diagnostic profiles based on statistically-driven multivariate classifications" (Kunina-Habenicht, Rupp, & Wilhelm, 2009, p. 64) of students according to the degree mastery on each of the requisite traits. Obtained information from profile scores can be used to tailor remediation for further instruction.

Technically speaking, cognitive diagnostic models (CDMs) are discrete and multidimensional latent variable models developed mainly for diagnosing students' mastery profiles on a set of sub/skills or attributes based on their observed item response patterns. According to Rupp and Templin (2008), CDMs are:

> "probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and non-compensatory ways to generate latent classes" (p. 226).

Like item response theory approach, CDMs are probabilistic models. They model the likelihood of a successful performance on a given item with respect to a number of latent traits or attributes. In unidimensional item response theory models, the probability of

generating a correct answer relies on a single latent trait, θ, so that those test takers with higher ability have a higher probability of success. However, CDMs explain a given student' performance level in terms of the probability of mastery of each attribute separately, or the probability of belonging to each latent class with a particular skill-mastery profile (Lee & Sawaki, 2009a).

CDMs are also confirmatory. Similar to confirmatory factor analysis, latent traits in CDMs are defined a priori through an incidence matrix called Q-matrix (Tatsuoka, 1983), which is considered as the loading structure of CDMs. It pinpoints a substantive hypothesis about the underlying response processes of students. The Q-matrix indicates the association between each item (rows) and its target cognitive subskills (columns) through a pattern of "1s" and "0s". If an item requires subskill k, qik=1; otherwise, qik=0. Additionally, Rupp and Templin (2008) state that another manifestation of confirmatory nature of CDMs is the priori specification of the way different attributes interact in the response process, that is, whether there exists a compensatory (disjunctive) or non-compensatory (conjunctive) relationship among the required attributes.

Furthermore, CDMs belong to multidimensional item response theory models. CDMs contain multiple latent traits in such a way that the successful performance on an item (or a task) requires the mastery of numerous sub-skills. Because each item is related to multiple attributes, CDMs have a complex loading structure. However, compared to multidimensional IRT and factor analysis (FA) in which latent traits are continuous, CDMs possess discrete or categorical latent variables.

With regard to assuming varying inter-skill relationships among the attributes, CDMs are classified into different categorizations. One common way is to differentiate between disjunctive/ conjunctive or compensatory/non-compensatory. According to compensatory models, the inadequacy of one attribute can be compensated for by the presence of other required attributes. Such models state that the mastery of more attributes does not increase the probability of success in a given test item. On the contrary, non-compensatory models assume that all the attributes are required to get an item right, that is, non-mastery of one attribute cannot be made up for by the mastery of other attributes. Lately, additive CDMs have been proposed as a new category of CDMs which assume that presence of any one of the attributes increases the probability of a correct response independent of the presence or absence of other attributes (Ma, de la Torre, & Sorrel, 2018).

Another important categorization of CDMs is specific vs. general. Specific CDMs are models which allow for only one type of relationship in the same test: conjunctive, disjunctive, and additive. On the other hand, general CDMs allow each item to select its own model that best fits it rather than imposing a specific model to all the items. de la Torre (2011) showed that several specific CDMs can be obtained from general models if appropriate constraints are applied in the parameterization of general models. For instance, the generalized deterministic inputs, noisy "and" gate (GDINA) (de la Torre, 2011), as a

general model, can be turned into DINA, DINO, ACDM, NC-RUM, and C-RUM by changing the link function into *log* and *logit* and setting the interaction effects to zero.

A wide array of CDMs with different theories or assumptions about the way of interaction between attributes (See Rupp & Templin, 2008; Ravand & Baghaei, 2019, for a review) have been proposed. The models include rule space methodology (RSM) (Tatsuoka, 1995), the attribute hierarchy method (AHM) (Leighton, Gierl, & Hunka, 2004), the higher-order DINA model (HO-DINA) (de la Torre, Douglas, & Jeffrey, 2004), the multi-strategy DINA (MS-DINA) (de la Torre, & Douglas, 2008), the DINO and NIDO models (Templin & Henson, 2006), the full noncompensatory reparameterized unifed model (full NC-RUM)/fusion model (Hartz, 2002; Roussos et al., 2007), the compensatory RUM (C-RUM) (de la Torre, 2011), the GDINA (de la Torre, 2011), the general diagnostic model (GDM) (von Davier, 2008; Xu & von Davier, 2008), the log-linear cognitive diagnosis model (LCDM) (Henson, Templin, & Willse, 2008), and the additive CDM (de la Torre, 2011). Most of these models have been applied in language assessment contexts on different language skills (Aryadoust, 2018; Buck & Tatsuoka, 1998; Buck, Tatsuoka, & Kostin, 1997; Buck et al., 1998; Chen & Chen, 2016; Effatpanah, Baghaei, & Boori, under review; Jang, 2009; Kasai, 1997; Kim, 2014; Lee & Sawaki, 2009a; Li, 2011; Li & Suen, 2013; Ranjbaran & Alavi, 2017; Ravand, 2016; Sawaki, Kim, & Gentile, 2009; Scott, 1998; Shahsavar, 2019; Sheehan, 1997; von Davier, 2008; Xie, 2016) and demonstrated to be useful for providing diagnostic feedback in service of instruction and learning (Nichols, 1994).

2. Literature Review

2.1 Cognitive Diagnostic Models

2.1.1 G-DINA

The G-DINA (de la Torre, 2011) is a general model which assumes both compensatory and non-compensatory relationships between attributes within the same test. In its saturated form, all possible interaction and main effects are considered. By imposing some limitations to main or interaction effects, several specific CDMs can be obtained from the model. Therefore, the probability of success for a test taker with a skill pattern α_{lj}^{*} is a function of the main effects and all the possible interaction effects among the k_j^{*} required skills for item *j* (de la Torre, 2011):

$$P(\alpha_{lj}^{*}) = \delta_{j0} + \sum_{k}^{k_j^{*}} \delta_{jk}\alpha_{lk} + \sum_{k' = k+1}^{k_j^{*}} \sum_{k=1}^{k_j^{*}-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots k_j^{*}} \prod_{k=1}^{k_j^{*}} \alpha_{lk}$$

where δ_{j0} is the intercept which represents the probability of a correct response when none of the required skills is present; δ_{jk} is the main effect due to attribute α_k; $\delta_{jkk'}$ is a first-order

interaction effect between α_k and $\alpha_{k'}$ which shows the change in the probability of a correct response due to the mastery of both α_k and $\alpha_{k'}$; $\delta_{j12\ldots k_j^*}$ is the highest-order interaction effect due to $\alpha_{1,\ \ldots,}\alpha_{k_j^*}$ which represents the probability of a correct response due to the mastery of all the required skills over and above the additive impact of all the main lower-order interaction effects (de la Torre, 2011).

2.1.2 DINA

The Deterministic Inputs, Noisy "and" Gate (DINA) (Haertel, 1989; Junker & Sijtsma, 2001) model is a non-compensatory or conjunctive model. It is regarded as the simplest and most restrictive CDMs which requires only two parameters for each item. Put simply, the DINA model partitions test takers into two deterministic latent groups (2^k) for each item. The first group includes examinees who have all required attributes to get an item right and the second group includes examinees who lack at least one of the main attributes measured by that item. In fact, lack of a single necessary attribute is the same as missing all required attributes. The probability of a correct response to an item by an examinee is:

$$P\,(X_{ij} = 1\,|\xi_{ij}) = \ (1 - \ s_i)^{\xi_{ij}}\, g_i^{1-\xi_{ij}}$$

where X_{ij} is a response for examinee *j* and item *i;* ξ_{ij} is a latent variable for examinee *j* and item *i*; s_i is the probability of a slip (an incorrect response to item *i* when all the required attributes have been mastered), g_i is the probability of a guess (a correct response to item *i* when none of the attributes have been mastered). According to de la Torre (2011), the DINA model can be derived from the G-DINA by setting all the parameters, e.g., main effects and lower order interaction effects, to zero: $g_{i1} = g_{i2} = 0$.

2.1.3 DINO

The Deterministic Input, Noisy, "or" Gate (DINO) (Templin & Henson, 2006) model is the compensatory analog to the DINA model. Like the DINA model, the DINO model has two parameters for each item. Examinees mastering at least one of the measured attributes for an item is expected to get the item right. Similar to the DINA model, the DINO model has the slipping and guessing parameters. The probability of a correct response for examinee *j* and item *i* can be expressed as:

$$P\,(X_{ij} = 1\,|\xi_{ij}) = \ (1 - \ S_i)^{\xi_{ij}}\, g_i^{1-\,\xi_{ij}}$$

where 1- s_i is the probability of not slipping for item *i*, and g_i is the probability of a guessing for item *i*. In terms of the parameters in the G-DINA, $\delta_{i0} = g_i$ and $1 - s_i' = \delta_{i0} + \delta_{ik}$ (de la Torre, 2011).

2.1.4 ACDM

 ISSN 2476-5880

Additive CDM (ACDM) (de la Torre, 2011) is a compensatory model which can be derived from the GDINA model by setting all the interaction effects to zero. The ACDM posits that the likelihood of producing a correct response increases by mastering each of the requisite attributes and lack of one attribute can be compensated for by the presence of other attributes. The ACDM has $K_j^* + 1$ parameters for item *j*. The item response function (IRF) for the ACDM is:

$$P\left(\alpha_{lj}^*\right) = \delta_{jo} + \sum_{k=1}^{k_j^*} \alpha_{jk} a_{lk}$$

2.1.5 NC-RUM

Non-compensatory Reparameterized Unified Model (NC-RUM) (de la Torre, 2011) or fusion model is a non-compensatory model which is similar to the ACDM in that all the interaction effects are equal to zero. Unlike the ACDM which has an identity link, the NC-RUM includes a log link function for estimation (de la Torre, 2011). The item response probability for an item required two attributes can be expressed as:

$$Log\ P(X_i = 1 \mid \alpha_1, \alpha_2) = \delta_{i0} + \delta_{i1}\alpha_1 + \delta_{i2}\alpha_2$$

2.1.6 C-RUM

Compensatory Reparameterized Unified Model (C-RUM) (Rupp et al., 2010) is the compensatory analog to the NC-RUM. Like the ACDM and NC-RUM, the C-RUM model can be derived from the GDINA by setting all the interaction effects to zero. However, this model is different from the NC-RUM in that it utilizes a logit link function instead of a log link function (de la Torre, 2011). The probability of a correct response for a two-attribute item is as follows:

$$Logit\ P(X_i = 1 \mid \alpha_1, \alpha_2) = \delta_{i0} + \delta_{i1}\alpha_1 + \delta_{i2}\alpha_2$$

2.2 Previous Applications of CDMs

As noted above, many researchers have applied a group of CDMs on different language skills, including reading (Buck, Tatsuoka, & Kostin, 1997; Chen & Chen, 2016; Jang, 2009; Kasai, 1997; Li, 2011; Li & Suen, 2013; Scott, 1998; Sawaki, Kim, & Gentile, 2009; Ranjbaran & Alavi, 2017; Ravand, 2016), listening (Aryadoust, 2018; Buck & Tatsuoka, 1998; Lee & Sawaki, 2009a; von Davier, 2008), and writing (Effatpanah, Baghaei, & Boori, under review; Kim, 2014; Shahsavar, 2019; Xie, 2016). In a pioneering study on the application of CDMs, Buck and Tatsuoka (1998) utilized the Rule Space Methodology to discover the underlying cognitive and linguistic attributes of listening comprehension. The results showed that fifteen prime attributes and fourteen interaction attributes explained 96%

of the variance of listening comprehension. They concluded that the rule space methodology can be used to accurately classify test takers into different latent knowledge states.

In another relevant study, Lee and Sawaki (2009a) conducted a ground-breaking multi-CDM study on the listening and reading sections of iBT TOEFL. They investigated the performance of three cognitive diagnostic models comprising the GDM, fusion model, and latent class analysis model (Yamamoto, 1982, 1990). The results of their analysis indicated that the three models perform similarly in terms of skill mastery probabilities, test takers skill mastery classification, and reliability of test takers classification.

Finally, in a recent study, Aryadoust (2018) compared the fit of five CDMs including the DINA, GDINA, DINO, HO-DINA, and RRUM to explore the underlying structure of the listening test of the Singapore-Cambridge General Certificate of Education (GCE) exam. He used only absolute and relative fit indices as criteria for comparing the models. The value of fit indices revealed that the RRUM has the optimal fit compared to the other models. The fit of the model was also supported by estimating classification consistency and accuracy. Further analysis showed that using world knowledge to make an inference is the most difficult attribute for test takers to master. He concluded that sub-skills of listening should be considered as non-compensatory in a sense that the lack of one attribute cannot be made up for by the presence of the other attributes.

In the present study, the research questions are as follows:

1- Which CDA model can better capture the diagnostic profile of the IELTS listening test more accurately compared to other CDMs?

2- What are the strengths and weaknesses of Iranian candidates in the listening section of the IELTS exam?

3. Methodology

3.1 Participants

The present study utilized the data Ghahramanlo et al. (2017) used for the application of the linear logistic test model (LLTM; Fischer, 1973). The data set includes scored responses of 310 participants to the listening section of the International English Language Testing System (IELTS). Of the total sample, there were 194 (62.9%) female and 116 (37.1%) male who ranged in age between 18 to 55 years (M= 25.32 years, SD= 5.65).

3.2 Instrumentation

The listening section of the IELTS exam was used in the study. The test composed of four sections, with 10 questions per section. The first two tasks concerned everyday social contexts and the last two tasks related to educational and training situations. In Task 1, students were required to listen to a woman being interviewed by a police officer about an incident she saw the previous evening. The test takers had to listen carefully to the woman as a victim and label the map based on the information she gives to the police officer. Also, they had to fill out a table associating with the physical appearance of thieves involved in the crime. There were two map labeling, four fill-in-the-gap and four multiple choice items. For one of the multiple choice questions, the participants were supposed to choose two correct answers.

In the second task, test takers were provided by a recorded message giving information about an English Hotel. Test takers had to answer questions relating to the location of the hotel, the facilities provided, and the price of accommodation in the hotel. The section comprised five multiple-choice and five fill-in-the gap items.

In task 3, examinees listened to three students talking about their study programs and a piece of advice given by one of the student. There were a multiple-choice and nine fill-in-the-gap items. Finally, in the last task, test takers listened to a talk by a university lecturer in Australia on a type of migratory bird. They answered to 10 fill-in-the-gap questions. After the completion of all the four tasks, test takers were given 10 minutes to correctly transfer their answers to a separate answer sheet. Reliability coefficients of the test were estimated using Cronbach alpha (α) (1951) analysis and a value of 0.91 was obtained which is highly acceptable.

Moreover, four experienced IELTS instructors were used for the stage of Q-matrix development. They were all non-native speakers of English, knowing Persian as their first language and English as their foreign language. Their sample included three IELTS instructors with more than 10 years of experience in teaching general English and IELTS and an educational supervisor with about 25 years of experience in teaching English and international high-stakes exams. The instructors held M.A. and Ph.D. degree in Teaching English as a Foreign Language (TEFL) and got band score 8 overall in IELTS exam. Their ages range from 32 to 53.

3.3 Q-matrix Specification

As a fundamental step in CDMs, an incidence matrix called Q-matrix (Tatsuoka, 1983) was developed to determine the conceptual relationship between a set of items and target

attributes. The quality of a cognitive diagnostic assessment is contingent upon the accurate specification of attributes underlying performance and their associations with test items. If a Q-matrix is misspecified, obtained information may result in invalid inferences (Rupp & Templin, 2008). Many methods have been suggested to define attributes involved in a test such as test specifications, content domain theories, analysis of item content, think-aloud protocol analysis of examinees' test-taking process, eye-tracking research, and the results obtained by the relevant research in the literature (Embretson, 1991; Leighton & Gierl, 2007; Leighton, Gierl, & Hunka, 2004). In the present study, four experienced IELTS instructors were considered as content experts to collectively indicate the major attributes required to perform correctly on each item. They were trained how to code the attributes measured by each item. A list of listening sub-skills introduced in various discussions about second language (L2) listening comprehension was given to the experts to specify what sub-skills are involved in the process of listening comprehension while listening the test items. The following list was identified for explaining the postulated attributes underlying the listening section of the IELTS:

- Making inferences (INF) (Tsui & Fullilove, 1998);
- Understanding paraphrases (PAR) (Wagner, 2004);
- Understanding detailed information (DET) (Sawaki et al., 2009);
- Understanding explicitly stated general and literal information (LIT) (Field, 2008);
- Comprehending vocabulary and syntax (VOG) (Aitkin, 1978; Shin, 2008; Wolfgram et al.,
- 2016);
- Keeping up with the pace of speakers (PAC) (Richards, 1983);
- Identifying prosodic patterns and speakers' attitudes and intentions (PPS) (Aitkin, 1978; Vandergrift, 2007).

Then, on the basis of the consensus among the experts on the item-subskill associations, an initial Q-matrix was developed. To empirically revise and validate the Q-matrix, the procedure suggested by de la Torre and Chiu (2016) using the "G-DIINA" package (Ma, de la Torre, & Sorrel, 2018) was utilized. In the first run of the analysis, some suggestions for the Q-matrix revision were provided. For example, it was suggested that understanding detailed information (DET) and understanding explicitly stated general and literal information (LIT) should be respectively involved for item 9 and 3. Admitting that statistical analysis should not be considered as the mere driving force for Q-matrix revision, the experts inspected the content of the item and agreed that these attributes are not necessary for the items. Also, for Items 34 and 36, it was suggested that making inference (INF) should be added to the Q-matrix. However, for items 21 and 24, the deletion of keeping up with the

ISSN 2476-5880

pace of speakers (PAC) was suggested. After several rounds of revisions and undertaking sensible modifications, the final Q-matrix presented in Table 1 was developed. Of the total items, nine of them were affiliated with INF, seven with PAR, twenty two with DET, fifteen with LIT, nine with VOG, twenty with PAC, and four with PPS. In Table 1, 1s indicate that the probability of producing a correct answer on each item is conditional on the mastery of the attributes whereas 0s show that the item does not need the sub-skills. As an illustration, in order for an examinee to get the item 5 right, he/she should have the mastery of INF, DET, and VOG.

Table 1: The Final Q-matrix

	INF	PAR	DET	LIT	VOG	PAC	PPS
1	1	0	1	0	0	0	1
2	0	0	1	0	0	0	1
3	1	0	0	0	0	1	0
4	1	1	0	0	0	0	0
5	1	0	1	0	1	0	0
6	0	0	0	1	0	0	0
7	0	0	0	1	1	0	0
8	0	0	0	1	0	0	0
9	0	0	0	1	0	0	0
10	0	0	1	0	1	0	0
11	0	1	1	0	1	0	0
12	0	1	0	1	1	0	0
13	0	1	0	1	0	1	0
14	0	0	0	1	0	0	0
15	0	0	0	1	0	0	0
16	0	0	0	1	0	0	0
17	0	0	0	1	1	0	0
18	0	0	0	1	0	0	0
19	0	0	0	1	0	0	0
20	0	0	1	0	0	1	0
21	0	0	0	1	0	1	0
22	0	0	0	1	0	1	1
23	1	0	0	0	0	1	1
24	0	0	1	0	1	0	0
25	0	0	1	0	1	1	0
26	0	0	1	0	0	1	0

 ISSN 2476-5880

27	0	0	0	1	1	0	0
28	0	0	1	0	0	1	0
29	0	0	1	0	0	1	0
30	0	0	1	0	0	1	0
31	1	1	1	0	0	1	0
32	0	1	1	0	0	1	0
33	0	1	1	0	0	1	0
34	1	0	1	0	0	1	0
35	1	0	1	0	0	1	0
36	1	0	1	0	0	1	0
37	0	0	1	0	0	0	0
38	0	0	1	0	0	1	0
39	0	0	1	0	0	1	0
40	0	0	1	0	0	1	0

4. Analyses and Results

The CDM package version 6.1-10 (Robitzch, Kiefer, George, & Uenlue, 2018) in the R statistical software (R core Team, 2013) was used to analyze the fit of six CDA models to the data including GDINA, DINA, DINO, ACDM, NC-RUM, and C-RUM. The CDM package generates different fit indices which can be used to determine the optimal model among the competing models (relative fit indices) and checking the fit of a model to the observed response data (absolute fit indices) (Rupp et al., 2010). To explore fit of the models at the test-level stage, a number of relative and absolute fit indices, as described below, were evaluated:

a) Akaike Information Criterion (AIC; Akaike, 1974). The AIC is a relative fit index employed to choose between non-nested models. The basic formula is defined as: AIC = -2LL + 2P where P is the number of parameters and LL is the log likelihood of the model.

b) Bayesian Information Criteria (BIC; Schwarz, 1978). Similar to AIC, the BIC is a relative fit index used to select between non-nested models. The basic formula is defined as BIC = -2LL + pln (N), where LL is the log likelihood of the model, P is the number of parameters in the model, and N is the sample size. Both AIC and BIC introduce a penalty for model complexity. Models with lower AIC and BICs are more preferable.

c) Mx2 (Chen & Thissen, 1997) is the test of global model fit which denotes the average of the X2 test statistics of independence for pairwise item response frequencies over all item pairs (Lei & Li, 2016). It is the mean difference between the model-predicted and observed response frequencies. As differences become larger, more evidence are gathered as

dependencies between the items. When CDM fits the data well, "the x2 test statistic is expected to be 0 within each latent class as the attribute profile of the respondents would perfectly predict the observed response patterns" (Rupp, Templin, & Henson, 2010, p. 269). Mx2 can be used for statistical significance (P-max X2) and a significant p-value indicates that the statistical independence of the item pair is violated and thus the model does not fit the data well (Hu et al., 2016).

d) The mean absolute difference for the item-pair correlations (MADcor) statistic (DiBello, Roussos, & Stout, 2006) is the average of absolute deviations between observed and predicted pairwise item correlations across all item pairs.

e) The average of absolute values of pairwise item covariance residuals (MADRESCOV; McDonald & Mok, 1995) is the mean discrepancy between matrices of observed and reproduced item correlations.

f) The standardized root mean square residual (SRMSR) is a fit index defined as the square root of the difference between the observed covariance (correlation) matrix and the model covariance matrix. Maydeu-Olivares (2013, p. 84) suggested models with SRMSR values below 0.05 as models with the "substantively negligible amount of misfit". However, Hu and Bentler (1999) suggested values below 0.08 as good fit.

4.1 Optimal Model Fit

Table 2 shows the relative and absolute fit statistics of the six models and the number of estimated parameters. As can be seen from the second column of the table, the GDINA model estimated 235 item parameters, DINA and DINO 109 parameters, and ACDM, C-RUM, and NC-RUM 155 parameters. It demonstrates that the DINA and DINO are parsimonious models and the GDINA is the most complicated model. As to the AIC, MX2, MADcor, SRMSR, and MADRES, the GDINA had the lowest values followed by the C-RUM, ACDM, NC-RUM, DINO, and DINA. However, with respect to BIC, the value of C-RUM was the lowest compared to the ACDM, DINO, DINA, NC-RUM, and GDINA. As BIC imposes a large penalty for more highly parameterized models, it is predictable for the GDINA model to have the worst value (Li, Hunter, & Lei, 2015). Overall, the C-RUM was the best fitting specific CDMs based on almost all indices. Therefore, the C-RUM is selected for further investigation to examine whether the model can accurately diagnose the performance of Iranian candidates in the Listening Sub-test of the IELTS. Previous studies have found that the C-RUM can better reflect the interaction of attributes in language assessment (Yi, 2012, 2017).

Table 2: Relative and Absolute Fit Indices

Models	Npars	AIC	BIC	MX2 (p)	MADcor	SRMSR	MADRES
GDINA	235	12359	13237	17.1 (0.028)	0.0483	0.0622	0.974
DINA	109	12686	13094	25.7 (0)	0.0652	0.0831	1.33
DINO	109	12669	13076	24.9 (0)	0.0646	0.0817	1.32
ACDM	155	12453	13032	23.7 (0)	0.0535	0.0697	1.08
NC-RUM	155	12579	13158	24.8 (0)	0.0591	0.076	1.31
C-RUM	155	12397	12976	19.9 (0.006)	0.0506	0.066	1.04

Table 3 provides further evidence for the fit of the C-RUM in terms of classification consistency P_c and classification accuracy P_a. As presented in Table 3, the classification accuracy (Pa) and consistency (Pc) for the whole latent class pattern is 0.80 and 0.71 respectively, indicating that the test possesses a 80% probability of accurately classifying a randomly selected respondent into his/her correct latent class from a single test administration. It also has a 71% probability of classifying a randomly selected respondent into the same category on different replications of the test. The other rows of the table show the consistency and accuracy of classifying examinees according to the mastery or non-mastery of each attribute. Similar to absolute fit statistics, there is not a definite criterion for Pa and Pc values. In the light of the results obtained by Cui et el. (2012), Wang et al (2015), and Johnson and Sinharay (2018), the values of accuracy and consistency are fairly high and acceptable in the current study.

Table 3: Classification Consistency P_c and Accuracy P_a

Classification Accuracy and Consistency	C-RUM
P_a	0.80
P_c	0.71
P_a INF	0.95
P_c INF	0.92
P_a PAR	0.90
P_c PAR	0.86
P_a DET	0.97
P_c DET	0.95
P_a LIT	0.95
P_c LIT	0.92

ISSN 2476-5880

P_a VOG	0.95
P_c VOG	0.91
P_a PAC	0.96
P_c PAC	0.93
P_a PPS	0.96
P_c PPS	0.94

4.2 C-RUM Analysis

As indicated in Table 4, of the seven sub-skills, making inference (INF) and comprehending vocabulary and syntax (VOG), mastered respectively by 27% and 45% of the examinees, were the most difficult attributes. Conversely, identifying prosodic patterns and speakers' attitudes and intentions (PPS) and understanding paraphrases (PAR) with 60% and 59% probabilities were the easiest sub-skills followed by understanding explicitly stated general and literal information (LIT), understanding detailed information (DET), and keeping up with the pace of speakers (PAC). It suggests that 60% of the students mastered PPS and 59% mastered PAR.

Table 4: Attribute Difficulty

Attributes	Attribute probability 1
INF	0.270
PAR	0.590
DET	0.498
LIT	0.564
VOG	0.450
PAC	0.456
PPS	0.600

As presented in Table 5, there are 128 viable latent classes (seven sub-skills with $2^7 = 128$ latent classes) with respect to the Q-matrix configuration. To save space, data for only a number of latent classes are shown. The table displays that the attribute profiles α_1= [00000] and $\alpha_{128} = [11111]$ were the most populated classes with 27% and 24% probabilities including approximately 85 and 74 persons respectively. The latent class 79 was the third populated sub-skill profile containing approximately 33 persons. The remaining profiles relate to respondents who mastered one of the attributes to six of the attributes.

ISSN 2476-5880

Table 5: Class Probabilities

Latent Class	Attribute Pattern	Class Probability	Class Expected Frequency
1	0000000	0.274	85.01
3	1000000	0.057	17.68
…	…	…	…
79	0111001	0.108	33.71
…	…	…	…
107	0101011	0.040	12.44
…	…	…	…
115	0100111	0.042	13.30
...	…	…	**...**
127	0111111	0.055	17.28
128	1111111	0.241	74.89

Table 6 shows, for space considerations, the C-RUM parameters for only the first two items. The first column gives the item number, the second column shows the required attributes for each item, the third column displays the attribute mastery patterns, and the last column represents the probability of a successful performance on each item with respect to the mastery of the required attributes by any given test item. As an illustration, successful performance on item 1 requires the presence of INF, DET, and PPS. Those test takers who have mastered none of the required attributes have only 15% probability of guessing to get the item right (e.g., item intercept). However, those test takers who have mastered INF have 34% chance to respond correctly to the item. In the same vein, those examinees who have mastered DET and PPS have 57% and 46% probability respectively. Also, respondents who have mastery of INF and DET have 95% probability to get the item right. By mastering the three attributes, the probability of responding correctly to the item increases to 98%.

Table 6: C-RUM Parameters

Item Number	Required Attributes	Mastery Patterns	Probability
I1	INF-DET-PPS	A000	0.15
I1	INF-DET-PPS	A100	0.34
I1	INF-DET-PPS	A010	0.57
I1	INF-DET-PPS	A001	0.46
I1	INF-DET-PPS	A110	0.95
I1	INF-DET-PPS	A101	0.78
I1	INF-DET-PPS	A011	0.94
I1	INF-DET-PPS	A111	0.98

 ISSN 2476-5880

I2	DET-PPS	A00	0.10
I2	DET-PPS	A10	0.37
I2	DET-PPS	A01	0.50
I2	DET-PPS	A11	0.98

Table 7 further provides the mastery probability of each examinee on any of the requisite attributes for a given test item or task. Due to the space limitation, the attributes mastery probability of only six randomly selected students are presented. The first column shows the student ID, followed by response pattern, attribute profile, the probability of belonging to this profile, and the attribute mastery probabilities. For instance, the probabilities that student 164 with the skill profile of [0101011] has mastered the attributes INF to PPS are 0.00, 0.85, 0.00, 0.74, 0.29, 0.99, and 0.99 respectively. In other words, there is a probability of 85% that he/she has mastered PAR and 0% probability for mastering INF and DET. The values above 0.50 shows a high confidence for the mastery status of different sub-skills for each student (Hu et al., 2016).

Table 7: Skill Mastery Probabilities

Test Takers	Response Pattern	Attribute Profile	P	INF	PAR	DET	LIT	VOG	PAC	PPS
4	11011011101111101110 01100101001000000001	0100111	0.34	0.27	0.98	0.04	0.43	0.63	0.51	0.99
64	11000010110111111111 00001001010000001000	0011000	0.55	0.00	9.05	0.62	0.88	0.31	0.29	0.31
111	11110111101011011010 11000001010000100000	0000111	0.64	0.00	0.22	0.00	0.13	0.96	0.98	0.98
164	11111111101111010011 01001001110000000000	0101011	0.70	0.00	0.85	0.00	0.74	0.29	0.99	0.99
243	11101111110011010000 00001011010010101001	0011000	0.93	0.00	0.04	0.99	0.99	0.01	0.00	0.04
301	11111110110111010111 01101011010111111011	1111111	0.91	0.91	0.98	0.99	0.99	0.99	0.99	0.99

Finally, Table 8 demonstrates the tetrachoric correlation among the attributes. The results show that there exists a moderate to strong correlation between the sub-skills. Overall, the values larger than 0.70 are considered as strong, 0.50 and 0.70 as moderate, and less than 0.50 as weak. Empirical studies showed that 0.50 is a logical value for correlation among attributes (e.g., Henson, Templin, & Douglas, 2007; Kunina-Habenicht, Rupp, & Wilhelm, 2012). As values indicate, there is a moderate correlation coefficients, which are italicized, between PAR and VOG (0.60), PAR and PAC (0.68), DET and PAC (0.57), and LIT and VOG (0.55). A high correlations is obvious among the most attributes.

Table 8: Tetrachoric Correlations between the Subskills

	INF	**PAR**	**DET**	**LIT**	**VOG**	**PAC**	**PPS**
INF	1						
PAR	0.98	1					
DET	0.88	0.76	1				
LIT	0.74	0.80	0.93	1			
VOG	0.99	*0.60*	0.74	*0.55*	1		
PAC	0.84	*0.68*	*0.57*	0.74	0.97	1	
PPS	0.81	0.89	0.81	0.91	0.89	0.98	1

5. Discussion

The present study aimed to serve two main purposes: (1) to select the best CDM for exploring how sub-skills underlying the listening section of the IELTS interact to produce a correct response and (2) to diagnose the performance of Iranian candidates in the Listening Sub-test of the IELTS exam. To answer the first research question, six cognitive diagnostic models, comprising the GDINA, DINO, ACDM, C-RUM, DINA, and NC-RUM, were compared at test-level. The results of relative and absolute fit indices showed that the GDINA model have a better performance among the rival models followed by the C-RUM, ACDM, NC-RUM, DINO, and DINA. The C-RUM as the best specific CDM was selected for further investigation. The better fit of C-RUM is starkly in line with Yi's (2012, 2017) studies who found that the C-RUM can better show the relationships among sub-skills involved in language assessment; however, it is in disagreement with Aryadoust (2018) who found the RRUM outperformed other CDMs for describing the underlying interaction among the listening sub-skills. Then, the fit of the C-RUM was further supported by analyzing the classification consistency and accuracy and tetrachoric correlations among the attributes. The results of the attribute-level and profile-level Pc and Pa indicated high and acceptable values for both pattern-level and subskill-level. Moreover, there were moderate to high correlations among the listening attributes. This can be considered as an evidence

for claiming that there exists a compensatory relationship among the L2 listening attributes. It is concordant with theories of listening comprehension which state that listening sub-skills are interdependent and complementary (Vandergrift & Goh, 2012). Harding et al. (2015) noted that "comprehension does not follow a strictly linear progression from the lower to the higher processing levels; rather, different levels may be operating concurrently, with breakdowns at one level compensated by "positive information" at another" (p.12).

Concerning the second research question, the analysis revealed that making inference (INF) and comprehending vocabulary and syntax (VOG) are the most difficult listening sub-skills. Also, the two "flat" skill mastery profiles, namely "non-master of all attributes" α_1= [0000000] and "master of all attributes" α_{128}= [1111111], were the most prevalent skill profiles. The existence of flat skill profiles can arise from either unidimensionality nature of the measured scale or the high correlations between the attributes (Lee & Sawaki, 2009a; Rupp et al., 2010). According to Lee and Sawaki (2009a),

> "… a CDA analysis may classify most of the examinees into flat profiles. This makes additional scores reported redundant, suggesting that reporting separate attribute scores provides little additional information over and above what a total score or overall proficiency score can offer. This can happen, for example, when a CDA is applied to a nondiagnostic test that was designed to be an essentially psychometrically unidimensional test for a target population (e.g., Luecht, Gierl, Tan, & Huff, 2006). When this happens, one can say that the utility of profile scoring is questionable from the psychometric point of view. This is a likely scenario in a domain such as language assessment where constructs are often found to be highly correlated among themselves" (p. 185).

As mentioned above, moderate to high correlations between the listening constituents were observed in the current study which can be considered as the reason for classifying most students into the flat skill profiles.

6. Conclusion

This study set out to find out what CDA model can reasonably reflect the underlying interaction among L2 listening comprehension and identify strengths and weaknesses of Iranian examinees in the listening section of the IELTS exam. The findings of the study showed that majority of the test takers are unable to have a successful performance on the test with respect to the interested attributes, especially in terms of making inference and understanding vocabulary and grammar. In this regard, it is suggested for IELTS instructors to attend more to these sub-skills in listening classes. By teaching and practicing difficult aspects of listening comprehension, students will have a better understanding of their deficiencies and adopt effective strategies to eliminate them.

As the process of all research faces some limitations, the present study might also suffer from some limitations which should be acknowledge and the conclusions drawn should be viewed within the constraints imposed on the study. The main limitation of the study was that a CDA approach was applied a non-diagnostic test which is problematic in terms of the validity of inferences about the test takers' skill mastery profiles (DiBello et al., 2006; Jang, 2009). An important area for further analysis is designing a true diagnostic test (Ravand & Baghaei, 2019) according to a CDA framework. However, retrofitting is worthwhile to determine the diagnostic capacity of existing achievement and proficiency tests before developing true diagnostic tests which need a big budget and a lot of time (Lee & Sawaki, 2009a).

In addition, the sample of the present study (N=310) was admittedly not very impressive for CDM application. Only a handful of studies have investigated the effect of sample size in the utilization of CDMs. These studies have shown that parameter recovery (Kunina-Habenicht, Rupp, & Wilhelm, 2012) and fit indices (Lei & Li, 2016) can be affected by sample size. In contrast, a few researchers believe that small sample size has more potential for recognizing the appropriate CDM (Choi et al., 2010; Hu et al., 2016; *Basokcu, 2014;* Maydeu-Olivares & Joe, 2014). Overall, cognitive diagnostic assessment has shown its promise for rich diagnostic information providing diagnostic information about the learning status of students. Consequently, more attention should be paid to designing and developing educational assessments in second/ foreign language contexts that are based on a CDM framework. Such an endeavor requires the cooperation of various experts from different fields of study (e.g., subject matter, measurement, pedagogy).

References

Aitken, K. G. (1978). Measuring listening comprehension in English as a second language. *TEAL Occasional Papers, Volume 2*. Vancouver, Canada: British Columbia Association of Teachers of English as an Additional Language.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723.

Alavi, S. M., Kaivanpanah, Sh., & Panahi Masjedlou, A. (2018). Validity of the listening Module of international English language testing system: Multiple sources of evidence. *Language Testing in Asia*, *8*(8), 1-17. doi.org/10.1186/s40468-018-0057-4

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An Individual differences approach. *Language Learning*, 62, 49-78.

doi.org/10.1111/j.1467- 9922.2012.00706.x

Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore–Cambridge general certificate of education O-level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International Journal of Listening*, 1-24. doi:10.1080/10904018.2018.1500915

Bas¸okc¸u, T. O. (2014). Classification accuracy effects of Q-Matrix validation and sample Size in DINA and G-DINA Models. *Journal of Education and Practice, 5*, 220–230.

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in Algebra using the Rule-space model. *Journal for Research in Mathematics Education, 24*(5), 442-459. doi:10.2307/749153

Britton, B. K., & Graesser, A. C. (2014). *Models of understanding text*. New York and London: Psychology Press.

Buck, G. (2001). *Assessing listening*. Cambridge, England: Cambridge University Press.

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. doi:10.1177/026553229801500201

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. doi:10.1111/0023-8333.00016

Buck, G. , Vanessen, T. , Tatsuoka, K. , Kostin, I. , Lutz, D. & Phelps, M. (1998). Development, Selection And Validation of a Set of Cognitive and Linguistic Attributes for the Sat I Verbal: Analogy Section. ETS Research Report Series, 1998: i-25. doi:10.1002/j.2333-8504.1998.tb01768.x

Carroll, J. B. (1972). Defining language comprehension. In R. O. Freedle & J. B. Carroll (Eds.), *Language comprehension and the acquisition of knowledge* (pp. 1–29). New York, NY: John Wiley and Sons.

Chapelle, C. A. (1994). CALL activities: Are they all the same? *System*, *22* (1), 33-45.

Chapelle, C. (1998). Multimedia call: Lessons to be learned from research on instructed SLA. Language Learning & Technology, 2 (1), 22-34. Retrieved April 20, 2006, from http://llt.msu.edu/vol2num1/article1/index.htm

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly, 13*(3), 218-230. doi:10.1080/15434303.2016.1210610

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289. doi:10.2307/1165285

Choi, H. J., Templin, J. L., Cohen, A. S., & Atwood, C. H. (2010, April). *The impact of*

Model misspecification on estimation accuracy in diagnostic classification models. paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:10.1007/bf02310555

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19-38. doi:10.1111/j.1745-3984.2011.00158.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199. doi:10.1007/s11336-011-9207-7

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273. doi:10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595. doi:10.1007/s11336-008-9063-2

de la Torre, J., & Douglas, & Jeffrey, A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353. doi:10.1007/bf02295640

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31A review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 979-1030): Elsevier.

Effatpanah, F., Baghaei, P., & Boori, A. A. (under review). Diagnosing EFL Learners' Writing Ability: A Diagnostic Classification Modeling Analysis. *Language Testing in Asia*.

Eom, M. (2008). Underlying factors of MELAB listening construct. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 77–94. Ann Arbor, MI: University of Michigan English Language Institute.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*(3), 495-515. doi:10.1007/bf02294487

Field, J. (2008). *Listening in the language classroom*. Cambridge, England: Cambridge University Press.

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension: An overview. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp.7–29).
Cambridge, England: Cambridge University Press.

Freedle, R. & Kostin, I. (1996). The prediction of TOEFL listening comprehension item difficulty for mini-talk passages: implications for construct validity. *TOEFL Research Report RR 96-29*. Princeton, NJ: Educational Testing Service.

Ghahramanlou, M., Zohoorian, Z., and Baghaei, P. (2017). Understanding the cognitive processes underlying performance in the IELTS listening comprehension test. International Journal of Language Testing. 7, 62–72.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321. Retrieved from http://www.jstor.org/stable/1434756

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*. Advance online publication. doi:10.1177/0265532214564505

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Henson, R. A., Templin, J. L., & Willse, J. T. (2008). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191. doi:10.1007/s11336-008-9089-5

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi:10.1080/10705519909540118

Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y.-H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing, 16*(2), 119-141. doi:10.1080/15305058.2015.1133627

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

IELTS (2017a). IELTS listening description. Retrieved from https://www.ielts.org/aboutthe-test/test-format-in-detail

IELTS (2017b). Test taker performance 2017. Retrieved from https://www.ielts.org/test-statistics/test-taker-performance

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. Language Testing, 26(1), 031-073. doi:10.1177/0265532208097336

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-aevel classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement, 55*(4), 635-664. doi:10.1111/jedm.12196

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272. doi:10.1177/01466210122032064

Kasai, M. (1997). Application of the rule space model to the reading comprehension section of the test of English as a foreign language (TOEFL) (Unpublished doctoral dissertation). University of Illinois at Urbana Champaign.

Kim, A.,Y. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227-258. doi:10.1177/0265532214558457

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*(2), 64-70. doi:10.1016/j.stueduc.2009.10.003

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*(1), 59-81. doi:10.1111/j.1745-3984.2011.00160.x

Lee, Y.-S., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: An empirical investigation. *Asia Pacific Education Review, 13*(2), 333-345. doi:10.1007/s12564-011-9196-3

Lee, Y.-W., & Sawaki, Y. (2009a). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172-189. doi:10.1080/15434300902985108

Lee, Y.-W., & Sawaki, Y. (2009b). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3), 239-263. doi:10.1080/15434300903079562

Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement, 40*(6), 405-417. doi:10.1177/0146621616647954

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205-237. Retrieved from http://www.jstor.org/stable/1435314

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3-16. doi:10.1111/j.1745-3992.2007.00090.x

Li, H. (2011). *Evaluating language group differences in the subskills of reading using a cognitive diagnostic modeling and differential skill functioning approach* (Doctoral dissertation). Pennsylvania State University, State College, PA.

Li, H., & Suen, H. K. (2013). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing, 30*(2), 273-298. doi:10.1177/0265532212459031

Li, H., Hunter, C. V., & Lei, P. W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33, 391–409.

doi.org/10.1177/0265532215590848

Liao, Y. (2007). Investigating the construct validity of the grammar and vocabulary section and the listening section of the ECCE: Lexico-grammatical ability as a predictor of L2 listening ability. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 5*, 37–78. Ann Arbor, MI: University of Michigan English Language Institute.

Ma, W., de la Torre, J., & Sorrel, M. (2018). *The Generalized DINA Model Framework* (R package version 2.0.8). Retrieved fro https://cran.rproject.org/web/packages/GDINA

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71-101. doi:10.1080/15366367.2013.831680

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305-328. doi:10.1080/00273171.2014.911075

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*(1), 23-40. doi:10.1207/s15327906mbr3001_2

Munby, J. (1978) *Communicative syllabus design*. Cambridge, England: Cambridge University Press.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575-603. doi:10.3102/00346543064004575

Nissan, S., DeVincenzi, F., & Tang, L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension*. (TOEFL Research Report RR 95-37). Princeton, NJ: Educational Testing Service.

Oakeshott-Taylor, J. (1977). Information redundancy, and listening comprehension. In R. Dirven (Ed.), *Hörverständnis im Fremdsprachenunterricht* [Listening comprehension in foreign language teaching]. Kronberg/Ts: Scriptor

Ranjbaran, F., & Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation, 55*, 167-179. doi:10.1016/j.stueduc.2017.10.007

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782-799. doi:10.1177/0734282915623053

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*. doi:10.1080/15305058.2019.1588278

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 17*, 219–239. doi:10.2307/3586651

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, C. (2018). *CDM: Cognitive diagnosis modeling* (Rpackage version 6.1-10). Retrieved from https://cran.r project.org/web/

packages/CDM/index.html

Rost, M. (2013). *Teaching and researching: Listening (2nd ed.)*. London: Routledge.

Rost, M. (2016). *Teaching and researching: Listening* (3rd ed.). London, UK: Longman.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275-318). Cambridge: Cambridge University Press.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219-262. doi:10.1080/15366360802490866

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY, US: Guilford Press.

Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*, 190–209. doi:10.1080/15434300902801917

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464. doi:10.1214/aos/1176344136

Scott, H. S. (1998). *Cognitive diagnostic perspectives of a second language reading test* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign, Urbana, IL.

Shahsavar, Z. (2019). Diagnosing English learners' writing skills: A cognitive diagnostic modeling study. *Cogent Education, 6*(1), 1-19. doi:10.1080/2331186X.2019.1608007

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*(4), 333-352. Retrieved from http://www.jstor.org/stable/1435113

Shin, S. (2008). Examining the construct validity of a web-based academic listening test: An investigation of the effects of response formats. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 95–129. Ann Arbor, MI: University of Michigan English Language Institute.

Snowling, M. J., & Hulme, C. (Eds.). (2005). *The science of reading: A handbook*. Oxford, UK: Blackwell.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354. Retrieved from http://www.jstor.org/stable/1434951

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In *Cognitively diagnostic assessment.* (pp. 327-359). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305. doi:10.1037/1082-989X.11.3.287

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics, 19,* 432–451. doi:10.1093/applin/19.4.432

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching, 40,* 191–210. doi:10.1017/S0261444807004338

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York, NY: Routledge.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287-307. doi:10.1348/000711007x193957

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. Span Fellow *Working Papers in Second or Foreign Language Assessment, 2*, 1–23. Ann Arbor, MI: University of Michigan English Language Institute.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement, 52*(4), 457-476. doi:10.1111/jedm.12096

Wolfgramm, C., Suter, N., & Göksel, E. (2016). Examining the role of concentration, vocabulary and self-concept in listening and reading comprehension. *International Journal of Listening*, 30, 25–46. doi:10.1080/ 10904018.2015.1065746

Xie, Q. (2016). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward? *Educational Psychology, 37*(1), 26-47. doi:10.1080/01443410.2016.1202900

Xu, X., & von Davier, M. (2008). FITTING THE STRUCTURED GENERAL DIAGNOSTIC MODEL TO NAEP DATA. *ETS Research Report Series*, i-18. doi:10.1002/j.2333-8504.2008.tb02113.x

Yamamoto, K. (1982). Hybrid model of IRT and latent class models. *ETS Research Report Series, 1982*(2), i-61. doi:10.1002/j.2333-8504.1982.tb01326.x

Yamamoto, K. (1990). *HYBILm: A computer program to estimate the HYBRID model*. Princeton, NJ: Educational Testing Service.

Yi, Y. (2012). Implementing a cognitive diagnostic assessment in an institutional test: A new Networking model in language testing and experiment with a new psychometric model and task type (Unpublished doctoral dissertation). University of Illinois at Urbana Champaign, Urbana-Champaign, IL.

Yi, Y. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. *Language Testing*, *34*(3), 337–355. doi.org/10.1177/0265532216646141