Tabaran Institute of Higher Education ISSN 2476-5880
International Journal of Language Testing
Special Issue, 2023: *Advanced Psychometric Methods in Language Testing*

A Cognitive Diagnostic Assessment Study of the Reading Comprehension Section of the Preliminary English Test (PET)

Aisha Mohammed[1]*, Abdul Kareem Shareef Dawood[2], Tawfeeq Alghazali[3], Qasim Khlaif Kadhim[4], Ahmed Abdulateef Sabti[5], Shaker Holh Sabit[6]

Abstract
Cognitive diagnostic models (CDMs) have received much interest within the field of language testing over the last decade due to their great potential to provide diagnostic feedback to all stakeholders and ultimately improve language teaching and learning. A large number of studies have demonstrated the application of CDMs on advanced large-scale English proficiency exams, such as IELTS, TOEFL, MELAB, and ECPE. However, too little attention has been paid to the utility of CDMs on elementary and intermediate high-stakes English exams. The current study aims to diagnose the reading ability of test takers in the B1 Preliminary test, previously known as the Preliminary English Test (PET), using the generalized deterministic input, noisy, "and" gate (G-DINA; de la Torre, 2011) model. The G-DINA is a general and saturated model which allows attributes to combine in both compensatory and non-compensatory relationships and each item to select the best model. To achieve the purpose of the study, an initial Q-matrix based on the theory of reading comprehension and the consensus of content experts was constructed and validated. Item responses of 435 test takers to the reading comprehension section of the PET were analyzed using the "G-DINA" package in R. The results of attribute profiles suggested that lexico-grammatical knowledge is the most difficult attribute, and making an inference is the easiest one.

Keywords: B1 Preliminary English test; reading attributes; G-DINA; Compensatory; non-compensatory

1. Introduction

Cognitive diagnostic models (CDMs), also referred to as diagnostic classification models (DCMs; Rupp & Templin, 2008), have been paid a great deal of attention in the field of language testing and assessment over the last decade due to their great promise for providing rich diagnostic information about weaknesses and strengths of test takers on a set of fine-grained attributes (Rupp et al., 2010). CDMs are probabilistic, confirmatory multidimensional latent variable models that have a simple or complex loading structure (Rupp & Templin, 2008). In contrast to classical test theory (CTT) and item response theory (IRT) as traditional psychometric frameworks which tend to locate persons on a unidimensional latent trait

[1]*College of Education/ Al-Farahidi University, Baghdad, Iraq; aishamohaisha7@gmail.com
[2]Department of English/ College of Arts/ Ahl Al Bayt University/ Kerbala, Iraq
[3]College of Media, Department of Journalism/ The Islamic University in Najaf, Najaf, Iraq
[4]English Language Department, Al-Mustaqbal University College, Babylon, Iraq
[5]Department of English, Al-Nisour University College, Baghdad, Iraq
[6]Scientific Research Center, Al-Ayen University, Thi-Qar, Iraq

continuum, CDMs offer formative feedback to all stakeholders based on mastery/non-mastery profiles of test takers which could be used to improve instruction and learning (DiBello et al., 1995). Different types of CDMs have been developed on the basis of different assumptions about the association between cognitive processes or attributes and students' responses to a set of given test items. Examples of CDMs include rule space methodology (RSM; Tatsuoka, 1983, 1995), the attribute hierarchy method (AHM; Leighton et al., 2004), the Deterministic Inputs, Noisy And Gate (DINA; Junker & Sijtsma, 2001), the Deterministic Input, Noisy, "or" Gate (DINO; Templin & Henson, 2006), the reduced reparametrized unified model (RRUM; Hartz, 2002), the additive CDM (ACDM; de la Torre, 2011), the linear logistic model (LLM; Maris, 1999), the generalized deterministic inputs, noisy and gate (G-DINA; de la Torre, 2011), the general diagnostic model (GDM; von Davier, 2008; Xu & von Davier, 2008), and the log-linear cognitive diagnosis model (LCDM; Henson et al., 2008).

Over the last few years, a large number of studies have utilized CDMs in educational and psychological measurement to diagnose mastery/non-mastery or absence/presence of several attributes. Generally, there are two approaches toward the use of CDMs in the literature. The first approach is to use CDMs to devise "true diagnostic" (Ravand & Baghaei, 2020, p.4) tests from the outset for diagnostic purposes. However, due to lack of diagnostic tests, majority of the CDM applications have been on the second approach, that is, to retrofit existing non-diagnostic tests in order to extract fine-grained diagnostic feedback beyond total scores. For instance, CDMs have already been applied on different sections of advanced large-scale English proficiency tests such as the IELTS (Effatpanah, 2019; Jang et al., 2020), TOEFL (Buck & Tatsuoka, 1998; Jang, 2005; Kasai, 1997; Kim, 2011; Lee & Sawaki, 2009a; Sawaki et al., 2009; von Davier, 2008; Yi, 2017a), MELAB (Li et al., 2015), ECPE (Templin & Bradshaw, 2014; Templin & Hoffman, 2013; Yi, 2017b), TOEIC (Buck et al., 1997), and LanguEdge field tests (Jang, 2009). Although these studies have provided valuable insights into the application of CDMs on non-diagnostic tests, there is a paucity of research on the use of CDMs on elementary and intermediate large-scale English proficiency tests such as B1 Preliminary.

B1 Preliminary, previously known as Cambridge English: Preliminary and the Preliminary English Test (PET), is an English language examination which is run by Cambridge Assessment English. B1 Preliminary is designed in two versions: (1) B1 Preliminary for schools, school-aged learners, and (2) B1 Preliminary for general and higher education adult learners. The test is one of the examinations in Cambridge English Qualifications designed to improve language skills of English learners. Each Cambridge English Qualification is appropriate for a particular level of the CEFR (e.g., the Common European Framework of Reference for Languages). As an intermediate-level qualification, the goal of the B1 Preliminary is to provide information about learners who have mastered the basics of English and now have practical language skills for everyday use. To the best knowledge of the authors, the cognitive processes underlying performance of the reading section of the test have not been yet fully explored, and that too little attention has been paid to the application of CDMs on intermediate-level examinations, especially B1 Preliminary. Therefore, this study aims to address two important concerns regarding the B1 Preliminary test: (1) the knowledge, processes, and (sub)skills test takers use to respond correctly to a set of test

items on the reading section of the test; and (2) the diagnosis and analysis of test takers' performance on the reading section of the B1 preliminary test. For this study, the following research questions are posed:

(1) What knowledge, processes, and (sub)skills B1 Preliminary test takers use to give a correct response to the items of the reading section of the test?
(2) What are the weaknesses and strengths of candidates in the reading section of the B1 Preliminary test?

2. Literature Review

2.1. Choosing CDMs for Reading Comprehension Tests

In the context of CDMs, the different elements of a specific (cognitive) domain are called attributes. In fact, attributes are any "procedures, skills, or knowledge a student must possess in order to successfully complete the target task" (Birenbaum et al., 1993, p. 443). In other words, attributes are viewed as domain-specific knowledge and (sub)skills required to show the mastery in a particular cognitive domain (Leighton & Gierl, 2007). For instance, considering reading comprehension as a general cognitive domain, lexico-grammatical knowledge, making an inference, skimming, and scanning are required attributes in order to comprehend a text and respond correctly to a set of reading comprehension items. The presence of multiple underlying cognitive attributes required for reading proficiency makes it a complicated process.

A critical consideration of studies using CDMs for analyzing reading comprehension is the selection of the most appropriate model. CDMs are typically classified into two categories: specific and general (Ravand & Baghaei, 2020). Specific CDMs are used in situations where a single type of relationship is feasible in the same test: non-compensatory, compensatory, and additive. In non-compensatory models, deficiency in one attribute cannot be completely made up for by other attributes. In other words, the mastery of all the required attributes for getting an item right is necessary. Compensatory CDMs, on the other hand, are models in which the mastery of any of the necessary attributes can make up for the lack of mastery of the other attributes. Additive CDMs are models in which the probability of success can be affected by the presence of any one of the attributes regardless of the absence or presence of other attributes. On the contrary, general CDMs do not assume any prespecified relationships among the attributes and allow multiple relationships in the same test.

In the literature of second/foreign language (L2) reading comprehension, there is a controversial view toward the interaction among reading attributes. Some researchers have supported the view that reading attributes are non-compensatory. As Hoover and Gough (1990) state, reading comprehension attributes should work together to allow a reader to understand a text, that is, deficiencies in any particular attribute cannot be made up for by the higher knowledge of other attributes. For that reason, a number of researchers have claimed that non-compensatory models can provide more accurate information compared to their compensatory counterparts (Li, 2011; Li, Hunter, & Lei, 2015; Roussos et al., 2007).

However, many researchers have argued that reading attributes are compensatory (e.g., Bernhardt, 2005; Coady, 1979; Goldsmith-Phillips, 1989; Stanovich, 1980; Stanovich & West, 1979, 1981). Coady (1979) argues that any deficiencies in one area can be

compensated for by the strength in other areas. Similarly, Stanovich (1980) introduced a compensatory-interactive model upon which "a deficit in any particular process will result in a greater reliance on other knowledge sources, regardless of their level in the processing hierarchy" (p. 32). In her compensatory model of second-language reading, Bernhardt (2005) explains that learners' first language (L1) reading ability and L2 knowledge make up for insufficiencies in each other during reading comprehension.

As there is mixed theoretical evidence with respect to the way reading attributes combine in a non-compensatory and compensatory relationship, several studies have shown that general CDMs such as the G-DINA model can better reflect the interactions among L2 reading attributes because they allow both compensatory and non-compensatory inter-attribute relationships (Du & Ma, 2021; Hemmati et al., 2016; Hemati & Baghaei, 2020; Lee & Sawaki, 2009a; Li, Hunter, & Lei, 2015; Ravand, 2016; Ravand & Robitzsch, 2018; Yi, 2017a).

2.2. G-DINA

The generalized deterministic inputs, noisy "and" gate (G-DINA; de la Torre, 2011) is considered as a general and saturated model which takes into account both non-compensatory and compensatory relationships among a set of attributes within a test. In its saturated form, the model considers all main and interaction effects and is equal to other general models for the purpose of cognitive diagnosis on the basis of alternative link functions (de la Torre, 2011). The G-DINA model classifies individuals into 2^{k_j}, where $K_j^* = \sum_{k=1}^{K} q_{jk}$ shows the number of required attributes for item *j*. It is assumed that each group has its own probability of success. For the G-DINA model (de la Torre, 2011), the probability of giving a correct response for an examinee with an attribute pattern α_{lj}^* is a function of the main effects and all the possible effects among the k_j^* required attributes for item *j*:

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

where δ_{j0} denotes the intercept for item *j*, which is defined as the probability of success when the required attributes are not present; δ_{jk} represents the main effect due to α_k; $\delta_{jkk'}$ is the interaction effect owing to α_k and $\alpha_{k'}$; and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$ (de la Torre, 2011).

With appropriate constraints on the parameterization of the G-DINA model, de la Torre (2011) showed that several constrained models which are special cases of the G-DINA model can be obtained. For example, by setting all the lower order interactions and the main effects, except for δ_{j0} and $\delta_{j12\ldots K_j^*}$, to zero, the DINA model can be derived. As a non-compensatory model, DINA assumes that all the required attributes should be present in order for an examinee to give a correct response to a given item. The compensatory counterpart of DINA is the DINO model which can be obtained from the G-DINA model by alternating sign which varies based on the order of interaction and constraining the magnitudes of the main and interactions to be identical to each other (de la Torre, 2011, pp. 182-183). As a compensatory model, the DINO assumes that the mastery of only one attribute increases the probability of success as the mastery of all the required attributes for the item. Additionally, several additive models with different link functions can be derived from the G-DINA model. The additive CDM (ACDM;

de la Torre, 2011) can be obtained if all the interaction effects are fixed to zero. The ACDM, as an additive and compensatory model, has $K_j^* + 1$ parameters for each item and assumes that the mastery of attribute α_k increases the probability of giving a correct response to item *j* in such a way that the contribution of the attribute is not dependent on the contribution of the other attributes (de la Torre, 2011). By setting all the interactions to zero, the LLM and RRUM can also be derived. Unlike the ACDM obtained by using an identity link, the LLM and RRUM can be obtained by using logit and log link, respectively. It should be noted that RRUM is a non-compensatory equivalent of the LLM (Hartz, 2002).

3. Method

3.1. Data

Data analyzed in this study includes item responses of 435 test takers to 25 items of the reading comprehension section of the B1 Preliminary English Test (PET). The test is made up four papers designed to measure test takers' English skills. The reading section generally consists of 35 questions in 6 parts, and there are various types of questions and texts. This section tends to measure the ability of test takers in reading and understanding the main points from magazines, newspapers, and signs. In the first part of the reading comprehension, there are five questions, and test takers have to read five real-world messages, notices, and other short texts to find the main point. The second section is composed of five questions in which test takers have to match five descriptions of people to eight short texts on a specific topic, assessing detailed comprehension. The third part includes ten questions, and candidates have to read a text to decide if each sentence is correct or incorrect. In the fourth part, there are five questions, and candidates have to read a long text for gist, detailed comprehension, and global meaning as well as writer's attitude and opinion. The last section is a multiple-choice cloze that contains ten questions. Test takers have to read a short text and choose the correct vocabulary items to complete gaps. Test takers are given 45 minutes to answer a set of given questions.

As previous studies showed that cloze items yield variance that is not explained by the reading comprehension dimension (Baghaei & Ravand, 2015; In'nami & Koizumi, 2009; Rauch & Hartig, 2010), ten items of multiple-choice cloze were removed from the study. Therefore, only 25 items of the test were analyzed. There were 156 males and 279 females who ranged in age between 19 and 39 (M=23.98, SD=4.210. The total score ranged from 0 to 24 with a mean of 12.01 and a standard deviation of 5.011.

3.2. Q-matrix Construction

The construction of a Q-matrix (Tatsuoka, 1983) is an important component of all CDMs because the diagnostic power of cognitive diagnostic assessment (CDA) relies heavily on the development and construction of a theoretically appropriate and empirically supported Q-matrix (Lee & Sawaki, 2009a, p. 169). A Q-matrix specifies the relationship between each item and its required attributes. This relationship is expressed by numbers 1 and 0. If the attribute is required by the item, number 1 is used; otherwise, number 0. Different ways have been proposed to identify attributes involved in a test, including verbal protocol analysis, eye-tracking research, test specifications, item content analysis, content domain theories, and relevant literature (Embretson, 1991; Leighton et al., 2004). As a non-diagnostic test was used in this study to extract diagnostic information, different methods were used to determine

attributes test takers should have mastered to correctly respond to a set of reading comprehension items. First, we consulted the literature on language proficiency and L2 reading comprehension models (Alderson & Lukmani, 1989; Hughes, 2003; Jang, 2009; Lumley, 1993; Munby, 1978) which consist of taxonomies of sub-skills and narrower domain of attributes for L2 reading comprehension. The second source for identifying reading attributes was brainstorming. As suggested by Lee and Sawaki (2009a), when no information on test specification and cognitive model of task performance is available, "brainstorming about possible attributes that elaborate on an existing test specification might serve as a good point of departure" (p. 176).

After reviewing the relevant literature on L2 reading comprehension, three certified English instructors as content experts with more than ten years of teaching experience in intermediate and advanced English courses (e.g., IELTS, TOEFL, FCE, CPE, and CAE) were invited to identify and brainstorm required attributes. Based on the review of the literature and experts' idea, a set of four attributes were specified: *lexico-grammatical knowledge (LGK)*, *main idea (MAI)*, *detailed information (DET)*, and *making an inference (INF)*. A 2-hour training session was also held to train experts on how to code the identified attributes measured by each item. They read each item and specified the attributes required to correctly answer reading comprehension items. Table 1a shows the initial Q-matrix.

In the next step, the initial Q-matrix was subjected to statistical procedure suggested by de la Torre and Chiu (2016) using the G-DINA package (Ma et al., 2022) in R to empirically validate the Q-matrix. This procedure depends on the G-DINA discrimination index (GDI) and looks for the simplest attribute specification of an item (de la Torre & Chiu, 2016). As the procedure identifies possible misspecifications and provides suggestions for revision of the Q-matrix, a few modifications were suggested. Except for one case for which the suggestion was to turn 1 into 0, in other cases, the suggestion was the insertion of 1 into the Q-matrix. All suggested modifications were carefully analyzed and only sensible suggestions were applied. To better understand the suggestions, mesa plot for each item was checked. Mesa plot visualizes a proportion of variance accounted for (PVAF) with respect to the maximum GDI of the item (de la Torre & Chiu, 2016; de la Torre & Ma, 2016). The mesa plot is a line graph on which the *x*-axis represents q-vectors or the potential item-attribute specifications, and the *y*-axis shows the corresponding PVAFs. In the mesa plot, the red dots are the original q-vectors. As noted by de la Torre and Ma (2016), the q-vector on the edge of the mesa is considered as the best attribute specification. Figure 1 illustrates mesa plots for items 11, and 12. Furthermore, Heatmap plot was checked to inspect the dependencies between item pairs using transformed correlations and log odds ratio. As can be seen in Figure 2, there is a dependency between items 24 and 25; however, applying the reasonable suggested modifications removed the dependency. The final Q-matrix is shown in Figure 1b.

Table 1

Initial and Final Q-matrices

(a) Initial Q-matrix

Items	LGK	MAI	DET	INF
1	1	1	0	0
2	1	1	0	0
3	1	1	0	0
4	1	1	0	0
5	1	1	0	0
6	1	0	1	0
7	1	0	1	0
8	1	0	1	0
9	1	0	1	0
10	1	0	1	0
11	0	0	0	1
12	0	0	0	1
13	0	0	0	1
14	1	0	0	1
15	1	0	0	1
16	0	0	0	1
17	0	0	0	1
18	1	0	0	1
19	0	0	0	1
20	1	0	0	1
21	0	1	0	0
22	0	0	1	0
23	0	0	1	0
24	1	1	0	1
25	0	0	0	1

(b) Final Q-matrix

Items	LGK	MAI	DET	INF
1	1	1	0	0
2	1	1	0	0
3	1	1	0	0
4	1	1	0	0
5	1	1	0	0
6	1	0	1	0
7	1	0	1	0
8	1	0	1	0
9	1	0	1	0
10	1	0	1	0
11	0	<u>1</u>	0	1
12	0	<u>1</u>	0	1
13	0	<u>1</u>	0	1
14	1	<u>1</u>	0	1
15	1	<u>1</u>	0	1
16	0	<u>1</u>	0	1
17	0	<u>1</u>	0	1
18	1	<u>1</u>	0	1
19	0	<u>1</u>	0	1
20	1	<u>1</u>	0	1
21	0	1	0	0
22	0	0	1	0
23	0	0	1	0
24	1	1	0	0*
25	0	0	<u>1</u>	1

Note: The underlined 1s represent the addition of attributes, and the asterisk * indicates the deletion of the attribute.

Figure 1

Mesa Plots for Items 11 and 12

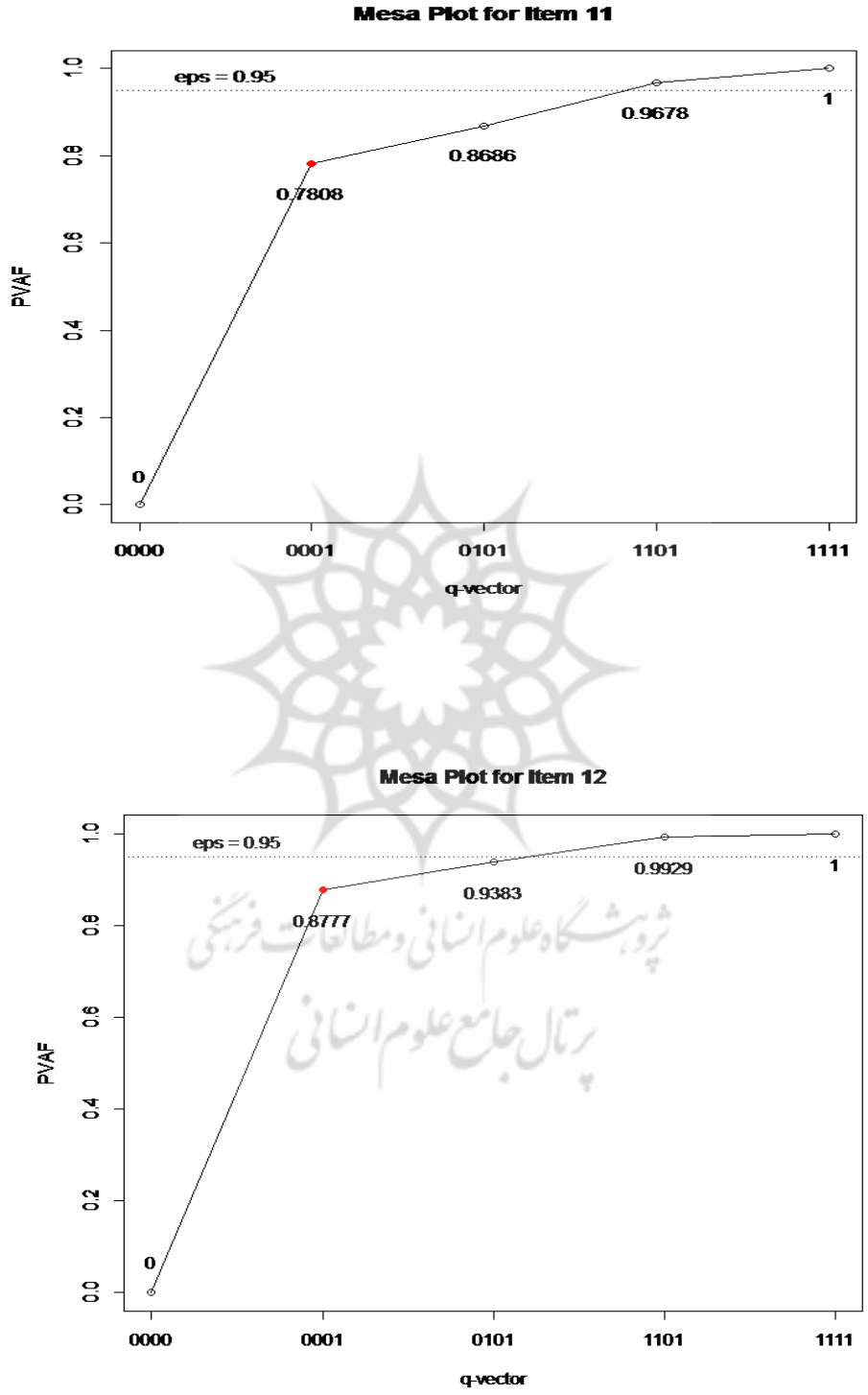
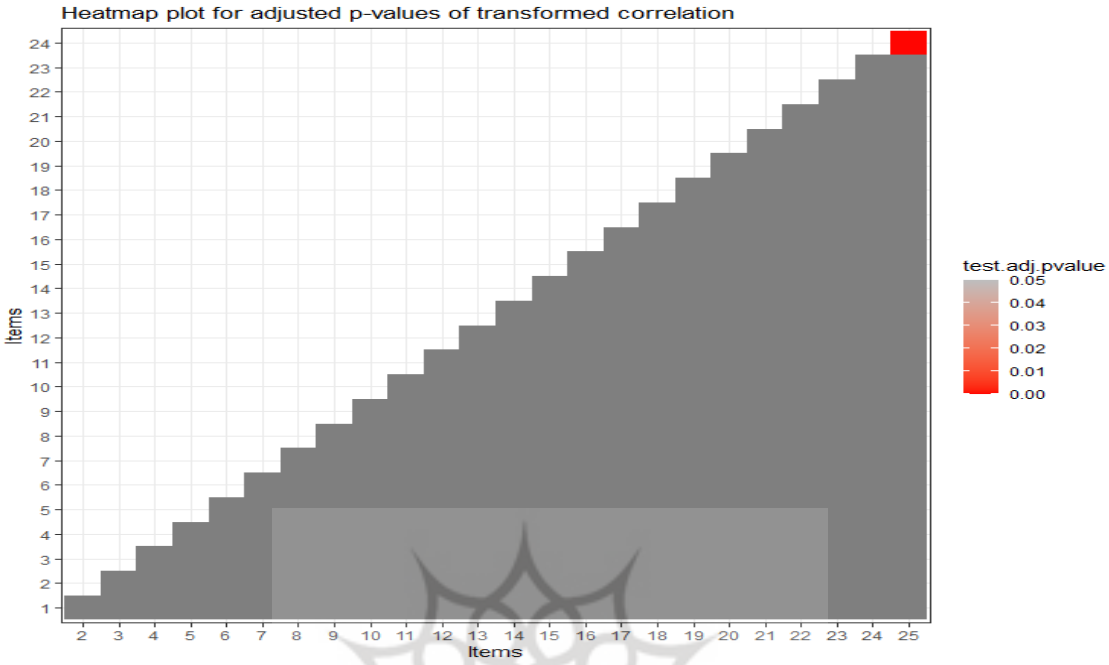


Figure 2

Heatmap Plots for the Initial and Final Q-matrices



(a) Heatmap Plot for the Initial Q-matrix

(b) Heatmap Plot for the Final Q-matrix

4. Results

4.1. Test-level Model Fit

For any statistical model, the fit of the model to the data should be firstly examined. The G-DINA package version 2.8.8 (Ma et al., 2022) in R (R Core Team, 2021) was used to examine the fit of the G-DINA to the data using marginal maximum likelihood estimation with EM algorithm. The package can produce two kinds of fit statistics to check the fit of a model (Rupp et al., 2010): (1) absolute fit statistics which are used to check the fit of a model to the data; and (2) relative fit statistics which are used to compare several models to select the best fitting one. In this study, different absolute fit indices were used to evaluate the fit of the G-DINA model to the observed response data. The absolute fit indices are as follows: (1) M_2 is considered as the mean difference between the model-predicted and observed response frequencies. Large values indicate that there are dependencies between the items. A significant *p*-value shows that the item independency is violated, and the model does not fit the data (Hu et al., 2016); (2) $RMSEA_2$ (the root mean square error of approximation fit index for M2) is a measure of difference between the observed covariance matrix and model-predicted covariance matrix for each degree of freedom (Chen, 2007, p. 467). This statistic ranges from 0 to 1. According to de la Torre (2011), values lower than 0.06 suggest good fit; (3) The standardized root mean squared residual (SRMSR) is the square root of the discrepancy between the observed covariance matrix and model-implied covariance matrix (Chen, 2007). As suggested by Hu and Bentler (1999), SRMSR is expected to be within the range of 0 and 0.08.

Table 2 summarizes the absolute fit results of the G-DINA model and the number of parameters. As demonstrated in the second column of Table 2, the G-DINA model estimated 125 item parameters. The value of M_2 was 231.84 with a non-significant *p*-value, suggesting good fit of the model to the data. The large value of M_2 could be due to the high number of items and small sample size. With regard to $RMSEA_2$, the value of the G-DINA model (e.g., 0.0191) and its upper and lower bounds were smaller than <0.06. Also, in relation to SRMSR, the value is within the ideal range of 0 and 0.08, e.g., 0.0451. Overall, the results of absolute fit statistics revealed that the G-DINA model has a satisfactory fit to the data.

Table 2
Absolute Fit Statistics

Fit Statistics	Npar	M_2 (p-value)	$RMSEA_2$	$RMSEA_2$ CI 1	$RMSEA_2$ CI 2	SRMSR
G-DINA	125	231.8458 (0.0608)	0.0191	0	0.0291	0.0451

Note: Npar = Number of parameters; CI: Confidence Intervals.

In addition to test-level absolute fit indices, three item-level fit statistics were also evaluated, including proportion correct (*p*), Log-odds ratio (*l*), and transformed correlations (*r*). According to Chen et al. (2013, p. 126), proportion correct (*p*) is the residual between the observed and predicted proportion correct of examinees' correct responses to a set of test items. Log-odds ratio (*l*) is the residual between the observed and predicted log-odds ratio of item pairs. And transformed correlations (*r*) is the residual between the predicted and observed

Fischer-transformed correlation of the item pairs. Smaller values indicate a better fit of the model to the data. Table 3 shows the absolute item-level fit for the G-DINA model. The results indicate that the model fits the data well. In terms of proportion correct, the *p*-value was lower than the critical Z-score (e.g., 4.17). The values of transformed correlations and log odds ratio were also acceptable because their adjusted *p*-value was smaller than 0.05, indicating good absolute fit of the model at the item level.

Table 3
Absolute Item-level Fit Indices

		mean[stats]	max[stats]	max[z.stats]	*p*-value	adj. *p*-value
G-DINA	**Proportion correct**	0.0009	0.0027	0.119	0.9052	1
	Transformed correlation	0.0358	0.1738	3.613	0.0003	0.0909
	Log odds ratio	0.1633	0.8347	3.723	0.0002	0.0591

Note: adj. *p*-value = adjusted *p*-value

Table 4 shows the prevalence of the four attributes. Making an inference (INF) was the easiest attribute for test takers because only 33% of test takers have not mastered this attribute. However, as can be seen, lexico-grammatical knowledge (LGK) has not been mastered by 50% of test takers, so LGK is the most difficult attribute, followed by main idea (MAI) and detailed information (DET).

Table 4
Attribute Prevalence

Attributes	Attribute Probability 0
Lexico-grammatical Knowledge (LGK)	0.5048
Main Idea (MAI)	0.4542
Detailed Information (DET)	0.4542
Making an Inference (INF)	0.3388

Based on the total number of attributes, CDMs classify test takers into 2^k latent classes. In this study, there are 16 latent classes with regard to the number of attributes or Q-matrix configuration ($2^4 = 16$). Table 5 demonstrates the proportion of the 16 attribute profiles for the G-DINA model, in which 1s show mastery of the required attributes, and 0s indicate non-mastery of the attributes. For example, attribute profile [1100] shows that the test taker has mastered the first and second attributes (e.g., lexico-grammatical knowledge (LGK) and main

idea (MAI)) and has not mastered the third and fourth attributes (e.g., detailed information (DET) and making an inference (INF)). As can be seen, the G-DINA model classified a large proportion of test takers into the last two latent classes, e.g., [0111] and [1111] with approximately 18% and 16%, respectively. Attribute profile of α_1 = [0000] and α_2 = [1000] had the third and fourth highest latent class probabilities of about 0.09 and 0.04, respectively, suggesting that approximately 9% of the examinees have mastered none of the required attributes, and about 4% have mastered only the first attribute (e.g, lexico-grammatical knowledge).

Table 5
Proportion of Attribute Mastery Profiles

Latent Class	Attribute Profile	G-DINA
1	0000	0.09834
2	1000	0.04036
3	0100	0.00000
4	0010	0.00000
5	0001	0.02596
6	1100	0.00000
7	1010	0.00000
8	1001	0.28960
9	0110	0.20010
10	0101	0.00000
11	0011	0.00000
12	1110	0.00000
13	1101	0.00000
14	1011	0.00000
15	0111	0.18040
16	1111	0.16530

Table 6 depicts the classification accuracy at both attribute- and test-level. According to Cui, Gierl, and Chang (2012), classification accuracy indicates to what extent individuals' "classification of latent classes based on the observed item response patterns agrees with their true latent classes" (p. 23). As can be seen in Table 6, values of classification accuracy at both attribute- and test-level for the G-DINA model were above 0.80, indicating a satisfactory classification rate (Cui et al., 2012; Effatpanah et al., 2019; Wang et al., 2015).

Table 6
Attribute- and Test-Level Accuracy

	Attribute-level Accuracy				**Test-level Accuracy**
G-DINA	**LGK**	**MAI**	**DET**	**INF**	
	0.8785	0.9540	0.9540	0.8835	0.8141

4.2. Item-level Model Fit

The second stage of the study was conducted to investigate whether the G-DINA model can be substituted by reduced CDMs (e.g., DINA, DINO, ACDM, LLM, and RRUM) without considerable loss in model data fit for each item using the Wald test. In this stage, each item can choose its best-fitting model. As stated by Ma, Iaconangelo, and de la Torre (2016), the Wald statistic is estimated for all constrained models for each multi-attribute item and then, if the null hypothesis is not confirmed ($p<.05$), the constrained model is rejected, and the G-DINA as a general model is selected. However, if several constrained models are retained and the DINA or DINO model is among the retained models, the DINA or DINO model with the larger *p*-value is chosen. But if the DINA or DINO are not retained, the constrained model with the largest *p*-value is selected. As can be seen in Table 7, among the 22 multi-attribute items, five items picked the DINA (e.g., 1, 5, 7, 9, and 19), five items the ACDM (e.g., 6, 8, 12, 24, and 25), five items the LLM (e.g., 2, 3, 4, 10, and 13), four items the G-DINA (e.g., 14, 15, 18, and 20), and three items the RRUM (e.g., 11, 16, and 17).

Table 7
Item-level Model Selection

Items	Attributes	Selected Model	p-value	Items	Attributes	Selected Model	p-value
1	LGK-MAI	DINA	0.2414	**12**	MAI-INF	ACDM	0.2988
2	LGK-MAI	LLM	0.5762	**13**	MAI-INF	LLM	0.8584
3	LGK-MAI	LLM	0.9964	**14**	LGK-MAI-INF	G-DINA	-
4	LGK-MAI	LLM	0.2768	**15**	LGK-MAI-INF	G-DINA	-
5	LGK-MAI	DINA	0.9054	**16**	MAI-INF	RRUM	0.8621
6	LGK-DET	ACDM	0.9319	**17**	MAI-INF	RRUM	0.5599
7	LGK-DET	DINA	0.8130	**18**	LGK-MAI-INF	G-DINA	-
8	LGK-DET	ACDM	0.7363	**19**	MAI-INF	DINA	0.6233
9	LGK-DET	DINA	0.7925	**20**	LGK-MAI-INF	G-DINA	-
10	LGK-DET	LLM	0.9937	**24**	LGK-MAI	ACDM	0.957
11	MAI-INF	RRUM	0.9889	**25**	DET-INF	ACDM	0.375

5. Discussion

This study was carried out to explore the cognitive processes of an intermediate large-scale English proficiency test, e.g., B1 Preliminary test, and diagnose reading ability of test takers based on their attribute profiles. As an important step in working with CDMs, a Q-matrix was developed. To identify attributes required to correctly answer reading comprehension items, we consulted the relevant literature on models of L2 reading comprehension and language ability. Three content experts were also recruited to brainstorm the required attributes.

In general, four attributes were specified, and an initial Q-matrix was developed. The identified attributes were Lexico-grammatical Knowledge (LGK), Main Idea (MAI), Detailed Information (DET), and Making an Inference (INF). After validating the Q-matrix through the procedure proposed by de la Torre and Chiu (2016), checking the mesa plot, and Heatmap plot, the data and the Q-matrix were subjected to analysis using the G-DINA as a general and saturated model. The results of absolute fit at both test- and item-level showed that the G-DINA model fits well to the data. The results of good fit of the G-DINA model were further supported by evaluating the classification accuracy. The analysis of classification accuracy showed that there are high values for attribute- and test-level accuracy, indicating the accurate classification of test takers into different latent classes.

The analysis of attribute profile patterns revealed that the test has adequate diagnostic power to distinguish between masters and non-masters across all the items. The finding of this study is in contrast to previous studies in which flat skill mastery profiles, e.g., "master of all attributes" and "non-master of all attributes", were reported as the most prevalent skill profiles for L2 reading comprehension, which could be due to the unidimensionality of the test or the existence of high correlations between attributes (Lee & Sawaki, 2009a). However, in this study, the test could classify test takers into different classes which is an indication of the significant performance difference between non-masters and masters.

The results of the study also showed that lexico-grammatical knowledge (KGK) is the most difficult attribute for test takers to master, and making an inference (INF) is the easiest attribute followed by main idea (MAI) and detailed information (DET). Previous studies showed that there is a hierarchy of difficulty of the L2 reading attributes (Baghaei & Ravand, 2015; Grabe & Stoller, 2002; Lumley, 1993; Ravand, 2016). According to Harding et al. (2015), "it is probably reasonable to accept that both first language and L2 reading involve a number of different 'levels' of ability" (p. 4). As Harding et al. (2015) argued, understanding main idea, making an inference, and understanding detailed information are considered as higher level attributes and lexico-grammatical knowledge is the lower level L2 reading comprehension attribute. Therefore, it is expected that making an inference to be difficult for test takers to master because it involves a higher level processing of the information in a given text (Grabe, 2009). However, the results of this study are in disagreement with previous studies because it was found that lexico-grammatical knowledge as a lower level attribute is the most difficult attribute. This could be due to the intermediate level of the test and test takers. The results of previous studies (see Ravand, 2016) were based on item responses of test takers obtained from advanced exams in which test takers are more likely to have more knowledge of vocabulary items and grammatical structures.

Finally, the item-level model fit was also checked to examine whether the G-DINA model can be replaced by the constrained models for items with more than one attribute. The results showed that the DINA is the best model for five items (e.g., 1, 5, 7, 9, and 19), the ACDM for five items (e.g., 6, 8, 12, 24, and 25), the LLM for five items (e.g., 2, 3, 4, 10, and 13), the G-DINA for four items (e.g., 14, 15, 18, and 20), and the RRUM for three items (e.g., 11, 16, and 17). The results of this study converge with previous CDM studies on L2 reading comprehension indicating that the interaction of L2 reading comprehension attributes can be

considered as a mixture of non-compensatory and compensatory (Lee & Sawaki, 2009a; Li et al., 2015; Ravand, 2016; Ravand & Robitzsch, 2018; Yi, 2012).

6. Conclusion

The main purpose of this study was to explore cognitive processes or attributes involved in answering the reading comprehension items of the B1 Preliminary test and identify the strengths and weaknesses of test takers in the reading section of the test. Based on the results of this study, a large proportion of test takers have failed to master some attributes that underlie the reading performance of the B1 Preliminary test. Therefore, both test takers and instructors in preparatory courses could use the results of this study to adopt some strategies and use effective materials to reduce or remove the deficiencies.

This study includes a few limitations which should be considered. First, the sample size of the current study was not very large enough to yield consistent results. Although many studies have shown that a small sample size has a great impact on fit statistics (Lei & Li, 2016) and parameter recovery of CDMs (Kunina-Habenicht et al., 2012), some researchers have argued that small sample sizes can better recognize the suitable CDM (Hu et al. 2016; Maydeu-Olivares & Joe, 2014). Furthermore, a non-diagnostic test was used in this study to extract diagnostic information. This approach has been shown to be troublesome with regard to the accuracy of inferences on the examinees' attribute mastery profiles (Jang, 2009). Lee and Sawaki (2009b) contend that

> "retrofitting efforts could serve as an important step in advancing cognitive diagnostic reading assessment research . . . it is worth examining the extent to which useful cognitive diagnostic information could be extracted from existing assessments before delving into an expensive, time consuming process of designing a new cognitive diagnostic test". (p. 174)

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5*, 253-270.

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using a linear logistic test model. *Learning and Individual Differences, 43*, 100-105. https://doi.org/10.1016/j.lindif.2015.09.001

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25*, 133-150. https://doi.org/10.1017/S0267190505000073

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. https://doi.org/10.2307/749153

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. https://doi.org/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. https://doi.org/10.1111/0023-8333.00016

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123-140 https://doi.org/10.1111/j.1745-3984.2012.00185.x

Coady, J. (1979). A psycholinguistic model of the ESL reader. In R. Mackay, B. Barkman, & R. R. Jordan (Eds.), *Reading in a second language* (pp. 5-12). Rowley, MA: Newbury House.

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. https://doi.org/10.1111/j.1745-3984.2011.00158.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199. https://doi.org/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. https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Ma, W. (2016, August). *Cognitive diagnosis modeling: A general framework approach and its implementation in R.* A Short Course at the Fourth Conference on Statistical Methods in Psychometrics, Columbia University, New York.

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-390). Hillsdale, NJ: Lawrence Erlbaum.

Du, W., & Ma, X. (2021). Probing what's behind the test score: Application of multi-CDM to diagnose EFL learners' reading performance. *Reading and Writing, 34*, 1441-1466. https://doi.org/10.1007/s11145-021-10124-x

Effatpanah, F. (2019). Application of cognitive diagnostic models to the listening section of the International English Language Testing System (IELTS). *International Journal of Language Testing, 9*(1), 1-28. URL: https://www.ijlt.ir/article_114295.html

Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia, 9*(12), 1-23. https://doi.org/10.1186/s40468-019-0090-y

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*(3), 495-515. https://doi.org/10.1007/BF02294487

Goldsmith-Phillips, J. (1989). Word and context in reading development: A test of the interactive-compensatory hypothesis. *Journal of Educational Psychology, 81*(3), 299-305. https://doi.org/10.1037/0022-0663.81.3.299

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*.

Cambridge, England: Cambridge University Press.

Grabe, W., & Stoller, F. (2002). *Teaching and research reading*. Harlow, UK: Longman.

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*, *32*(3), 317-336. https://doi.org/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.

Hemati, S. J., & Baghaei, P. (2020). A cognitive diagnostic modeling analysis of the English reading comprehension section of the Iranian National University Entrance Examination. *International Journal of Language Testing*, *10*(1), 11-32. URL:https://www.ijlt.ir/article_114278_f909e5fbb63a5e1bb33d0da245ee9e53.pdf

Hemmati, S. J., Baghaei, P., & Bemani, M. (2016). Cognitive diagnostic modeling of L2 reading comprehension ability: Providing feedback on the reading performance of Iranian candidates for the University Entrance Examination. *International Journal of Language Testing, 6*(2), 92-100. URL:https://www.ijlt.ir/article_114432_d0c792401cf6619358a88c4591172ab6.pdf

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), 191-210. https://doi.org/10.1007/s11336-008-9089-5

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*, 127-160. https://doi.org/10.1007/BF00401799

Hu, L., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55. https://doi.org/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. https://doi.org/10.1080/15305058.2015.1133627

Hughes, A. (2003). *Testing for language teachers* (2nd Ed.). New York: Cambridge University Press.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219-244. https://doi.org/10.1177/0265532208101006

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

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), 31-73. https://doi.org/10.1177/0265532208097336

Jang, E. E., Kim, H., Vincett, M., Barron, C., and Russell, B. (2019). Improving IELTS reading test score interpretations and utilisation through cognitive diagnosis model-based skill profiling. *IELTS Research Reports Online Series, No. 2. British Council, Cambridge Assessment English and IDP: IELTS Australia.* Available at

https://www.ielts.org/teaching-and-research/research-reports

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. https://doi.org/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, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing*, *28*(4), 509-541. https://doi.org/10.1177/0265532211400860

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. https://doi.org/10.1111/j.1745-3984.2011.00160.x

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

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

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. https://doi.org/10.1177/0146621616647954

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. https://doi.org/10.1111/j.1745-3992.2007.00090.x

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

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *9*, 17-46. Retrieved from https://michiganassessment.org/research/research-database

Li, H., Hunter, C. V., & Lei, P. W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, *33*(3), 391-409. https://doi.org/10.1177/0265532215590848

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, *10*(3), 211-234. https://doi.org/10.1177/026553229301000302

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*(3), 200-217 https://doi.org/10.1177/0146621615621717

Ma, W., de la Torre, J., Sorrel, M., & Jiang, Zh. (2022). *G-DINA: The generalized DINA*

model framework. R package version 2.8.8.
Retrieved from https://CRAN.R-project.org/package=G-DINA

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

Mirzaei, A., Heidari Vincheh, M., & Hashemian, M. (2020). Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context. *Studies in Educational Evaluation, 64*, 1-10. https://doi.org/10.1016/j.stueduc.2019.100817

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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354-379.
URL:https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2010_20101218/02_Rauch.pdf

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

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing, 20*(1), 24-56. https://doi.org/10.1080/15305058.2019.1588278

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: A study of reading comprehension. *Educational Psychology, 38*(10), 1255-1277. https://doi.org/10.1080/01443410.2018.1489524

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*(4), 293-311. https://doi.org/10.1111/j.1745-3984.2007.00040.x

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. https://doi.org/10.1080/15366360802490866

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

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*(3), 190-209. https://doi.org/10.1080/15434300902801917

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*(1), 92-111. https://doi.org/10.2307/747348

Stanovich, K. E., & West, R. F. (1979). Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Memory & Cognition, 7*, 77-85.

https://doi.org/10.3758/BF03197588

Stanovich, K. E., & West, R. F. (1981). The effect of sentence context on ongoing word recognition: Tests of a two-process theory. *Journal of Experimental Psychology*: *Human Perception and Performance*, *7*(3), 658-672. https://doi.org/10.1037/0096-1523.7.3.658

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconception based on item response theory. *Journal of Education Measurement*, *20*(4), 345-354 https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*, 317-339. https://doi.org/10.1007/s11336-013-9362-0

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

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37-50. https://doi.org/10.1111/emip.12010

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287-307. https://doi.org/10.1348/000711007X193957

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. https://doi.org/10.1111/jedm.12096

Xu, X., & von Davier, M. (2008). Fitting the structured general diagnostic model to NAEP data. *ETS Research Report Series*, *2008*(1), i-18. doi:10.1002/j.2333-8504.2008.tb02113.x

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. (2017a). 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. https://doi.org/10.1177/0265532216646141

Yi, Y. (2017b). In search of optimal cognitive diagnostic model(s) for ESL grammar test Data. *Applied Measurement in Education*, *30*(2), 82-101. https://doi.org/10.1080/08957347.2017.1283314