Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 14, No. 1; 2024, March

ISSN 2476-5880
https://www.ijlt.ir

A Cognitive Diagnostic Modeling Analysis of the Reading Comprehension Section of an Iranian High-Stakes Language Proficiency Test

Ali Akbar Boori[1]*, Mohammad Ghazanfari[2]*, Behzad Ghonsooly[3], Purya Baghaei[4]

ARTICLE INFO

Article History:
Received: May 2023
Accepted: June 2023

KEYWORDS
Additive
Attributes
CDMs
Compensatory
G-DINA
Non-compensatory
Reading Comprehension

ABSTRACT

The purpose of this study was to compare the functioning of five restrictive CDMs, including DINA, DINO, *A*-CDM, LLM, and RRUM, against the G-DINA model to identify the best-fitting CDM which can better explain the interaction underlying the attributes of the reading comprehension section of an Iranian high-stakes language proficiency test. To this end, the performance of 1152 examinees to the reading section of the test was examined. The six CDMs were initially compared in terms of relative and absolute fit statistics at test-level to choose the best model. It was found that the G-DINA model outperformed compared to the restrictive models; thus, it was chosen for the second phase of the study. Concerning the second purpose of the study, the G-DINA was used to identify the strong and weak points of the reading proficiency of the test takers. The results revealed that making an inference and vocabulary are the hardest attributes for examinees of the test, and understanding the specific information is the easiest attribute. Finally, the models were also compared at item-level. The presence of a combination of L2 reading attributes was found.

1. Introduction

Cognitive diagnostic assessment (CDA), as a new kind of educational assessment, has been introduced to assess detailed or specific knowledge structure and provide formative diagnostic feedback about students' strong and weak points of the reading proficiency of the test takers to improve further learning and teaching (Rupp & Templin, 2008). CDA is the outcome of combining cognitive psychology and educational measurement for understanding the learning status of examinees. Messick (1989) emphasized the importance of understanding test performance in terms of the mental or cognitive processes examinees adopt to get an item right (e.g., substantive approach), which is the main feature of construct validity. Messick (1989) argued that

> In the substantive approach, items are included in the original pool on the basis of judged relevance to a broadly defined domain but are selected for the test on the basis of empirical response consistencies. The substantive component of construct validity is the ability of the construct theory to account for the resultant test content. . . .the internal structure and substance of the test can be addressed more directly by means of causal modeling of item or task performance. This approach to construct representation attempts to identify the theoretical mechanisms that underlie task performance, primarily by decomposing the task into requisite component processes (Embretson, 1983). Being firmly grounded in the cognitive psychology of information processing, construct representation refers to the relative dependence of task responses on the

[1] Ferdowsi University of Mashhad, Email: aaboori@gmail.com
[2]* Ferdowsi University of Mashhad, Email: mghazanfari@ferdowsi.um.ac.ir
[3] Ferdowsi University of Mashhad, Email: ghonsooly@um.ac.ir
[4] Islamic Azad University, Mashhad Branch, Mashhad, Email: puryabaghaei@gmail.com

Cite this paper as: Boori, A., Ghazanfari, M., Ghonsooly, B., & Baghaei, P. (2024). A Cognitive Diagnostic Modeling analysis of the reading comprehension section of an Iranian high-stakes language proficiency test. *International Journal of Language Testing*, *14*(1), 17–33. https://doi.org/10.22034/IJLT.2023.399561.1256

processes, strategies, and knowledge (including self-knowledge) that are implicated in test performance. (pp. 42-45)

Because CDAs are intrinsically diagnostic, advanced statistical models, the so-called cognitive diagnostic models (CDMs; Rupp & Templin, 2008) are utilized to measure the extent to which examinees have mastered a set of sub-skills required for successful performance on a set of given test items. Rupp and Templin (2008, p. 226) define CDMs as "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". In CDMs, tasks and/or items are broken down into various operations, strategies, and knowledge examinees need to give a correct response to given items (Embretson, 1983). Diverse constituents of a cognitive domain are known as attributes in CDM literature, also known as attributes, sub-skills, skills, processes, abilities, strategies, and knowledge. Birenbaum et al. (1993) define attributes as any "procedures, skills, or knowledge a student must possess in order to successfully complete the target task" (p. 443). For instance, reading is a general cognitive domain calling for a number of attributes including vocabulary, grammar, understanding explicit information, identifying specific information, making an inference, and so on. Text comprehension and successful performance on several test items require readers to have mastered these attributes. This feature allows CDMs to generate multidimensional diagnostic profiles according to the mastery/non-mastery of each necessary attribute (Helm et al., 2022).

In the last few decades, numerous CDMs have been proposed. The models include the the Deterministic Inputs, Noisy "and" Gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001), the Deterministic Inputs, Noisy "or" Gate (DINO; Templin & Henson, 2006), the reduced reparameterized unified model (RRUM or fusion model; Hartz, 2002), the linear logistic model (LLM; Maris, 1999), the additive CDM (A-CDM; de la Torre, 2011), the Generalized Deterministic, Inputs, Noisy "and" Gate (G-DINA; de la Torre, 2011), the general diagnostic model (GDM; von Davier, 2008), and the log-linear cognitive diagnosis model (LCDM; Henson et al., 2008). Particularly relevant to this study, a large number of studies have been conducted to apply different CDMs on reading comprehension skill (e.g., Buck et al., 1997; Chen & Chen, 2016; Du & Ma, 2021; Hemati & Baghaei, 2020; Jang, 2009; Jang et al., 2019; Javidanmehr & Anani Sarab, 2019; Kasai, 1997; Mehrazmay et al., 2021; Mirzaei et al., 2020; Ranjbaran & Alavi, 2017; Ravand, 2015; Sawaki et al., 2009; Wang & Gierl, 2011). The results of these studies showed that CDMs can yield reliable and valid diagnostic information about the reading ability of examinees.

A major issue in the application of CDMs to L2 reading is that of selecting the most appropriate CDM. In fact, a critical decision for researchers and practitioners is that L2 reading comprehension attributes interact in a compensatory, non-compensatory, or additive manner, or even all types of relationships are at work. As the primary purpose of CDMs is to classify individuals into different latent classes based on their observed response patterns, selecting the inappropriate model would result in inaccurate classification and misleading feedback (Lee & Sawaki, 2009a), that is, there should be a match between the assumptions of the models and the way the attributes underlying a skill interact.

This study aimed to compare the performance of five constrained CDMs, including DINA, DINO, A-CDM, LLM, and RRUM, against the G-DINA to explore the most optimal CDM which can better account for the underlying interaction between attributes of the reading comprehension section of the Islamic Azad University English Proficiency Test (IAUEPT), as an Iranian high-stakes test devised for measuring language ability of candidates who tend to pursue their studies at Ph.D. level. Then, the best-fitting model is used to determine the strong and weak points of the reading proficiency of the test takers. To fulfill the aim of this study, the research questions were posed as follows:

Q1. How do the G-DINA, DINA, DINO, A-CDM, LLM, and RRUM fit the reading comprehension section of the IAUEPT at test- and item-level?
Q2. What is the most optimal model for the reading comprehension section of the IAUEPT?
Q3. What are the strong and weak points of the reading proficiency of the test takers in the reading comprehension section of the IAUEPT?
Q4. To what extent, can the items of the IAUEPT reading comprehension section provide diagnostically useful information?
Q5. To what extent does the diagnosis approach provide accurate skill mastery classification?

2. Review of Literature

2.1 The Choice of CDMs for Reading Tests

CDMs have been categorized into two general forms (Ravand & Baghaei, 2019). Specific CDMs refer to models which assume only one sort of relationship within the same test: compensatory, non-compensatory, and additive. In compensatory models such as DINO, the assumption is that the absence or a low level of competence on attributes can be compensated for by the presence or a high level of competence on the other attributes. However, non-compensatory models assume that the absence or a low level of competence on one attribute cannot be compensated for by the presence or a high level of competence on the other attributes. In effect, individuals need to have mastered all the required attributes to give a correct answer to a test item. A well-known example of a non-compensatory model is DINA. Additive models further assume that the presence of each attribute affects the probability of getting an item right irrespective of the presence or absence of the other required attributes. It must be noted that LLM is both an additive and a compensatory model whereas the RRUM is both an additive and a non-compensatory model. In contrast, general CDMs permit different types of relationships among the attributes within the same test. In fact, each item can select the model that has the best fit. The G-DINA, GDM, and LCDM are examples of general CDMs.

Reading comprehension in a second/foreign language (L2) is a complex cognitive process which requires decoding and linguistic knowledge (Gottardo & Mueller, 2009; Grabe, 2009; Koda, 2005; Shiotsu & Weir, 2007; Zhang, 2012). Over the last few decades, numerous researchers have developed various reading comprehension models (Bernhardt, 2005; Gough & Tunmer, 1986; Kintsch & Rawson, 2005; Koda, 2005; McNeil, 2012; Sadoski & Paivio, 2007; Stanovich, 1980; Stanovich & West, 1981; Weir et al., 2000) and taxonomies (Hughes, 2003; Jang, 2009; Munby, 1978) to explain reading comprehension and its underlying attributes.

A major issue on studying L2 reading attributes is the type of interactions or relationships that exist among the attributes. Researchers have taken different views toward the relationships between reading attributes, that is, whether reading attributes interact in a non-compensatory or compensatory manner. In the literature, there are researchers who have maintained that there exists a compensatory interplay between L2 reading attributes (e.g. Stanovich & West, 1981). For example, Stanovich (1980) contends that a lack of competence in one area or a particular process can be offset by strength in another area or other processes, suggesting the compensatory nature of reading comprehension. The interactive model of reading introduced by Stanovic and West (1981) further stated that strengths in top-down processes can compensate for any deficits in the bottom-up processes engaged in reading comprehension. Using data from children in different grades (e.g., second, fourth, and sixth grades), Goldsmith-Philips (1989) conducted a study to test the model of Stanovic (1980). The results supported the interactive-compensatory nature of reading comprehension attributes. Bernhardt (2005) proposed a model of L2 reading assuming that first language (L1) and L2 reading proficiency combine in a compensatory manner. Usó-Juan (2006) also indicated that English language ability and discipline-related knowledge of a learner can compensate reciprocally. For that reason, several researchers have argued that compensatory CDMs can better reflect the interaction of reading attributes (Li et al., 2015; Yi, 2012).

On the other hand, some researchers have devoted particular attention to the non-compensatory nature of reading comprehension (Sadoski & Paivio, 2007). In their Simple View of Reading, Gough and Tunmer (1986) argued that reading comprehension is the outcome of (text) decoding and linguistic comprehension, and both components have equal importance, suggesting the non-compensatory nature of reading comprehension. In much the same vein, Hoover and Gough (1990) maintained that reading components, such as decoding and linguistic knowledge, should work together to achieve successful reading comprehension. A lack of competence in one component cannot be compensated for by the competence in another component. According to the dual-coding model, Sadoski and Paivio (2007) showed that both verbal (e.g., linguistic comprehension) and visual processes (e.g., imagery and matching information) work in unison to help readers to comprehend a text, indicating the non-compensatory nature of reading comprehension. Many researchers, therefore, have preferred non-compensatory CDMs over compensatory models for analyzing reading comprehension (Buck et al., 1997; Kasai, 1997; Kim, 2015; Jang, 2009; Li, 2011; Li & Suen, 2013; Roussos et al., 2007, to name a few).

Previous studies have witnessed a mixed and controversial view toward the relationship between reading attributes. As a result, several studies have indicated that reading comprehension attributes combine in both non-compensatory and compensatory manners (Jang, 2009). In this way, general CDMs like the G-DINA, GDM, and LCDM can better capture the possible interaction between reading comprehension attributes (Du & Ma, 2021; Ravand, 2015; Ravand & Robitzsch, 2018; Yi, 2017).

2.2 General and Specific CDMs Used in this Study

2.2.1 GDINA

As a saturated and general model, G-DINA (de la Torre, 2011) is considered as the generalization of the DINA model. The model consists of all possible interaction and main effects. In the DINA model, test takers are partitioned into two groups, but the G-DINA model partitions examinees into $2^{k_j^*}$ classes, where k_j^* is the number of necessary attributes for item *j*. In effect, each group has its own probability of success. The probability of correctly responding to an item for an examinee 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:*

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

where δ_{j0} is the intercept for item *j* (e.g., the probability of a correct response when none of the required skills is present); δ_{jk} is the main effect due to a single attribute α_k, showing the change in the probability of success as a result of mastering a single attribute (i.e., a_k); $\delta_{jkk'}$ is the (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 shows the probability of a correct response due to the mastery of all the required skills that is above and over the additive impact of all the main lower-order interaction effects (de la Torre, 2011, p. 181).

2.2.2 DINA

DINA (Haertel, 1989; Junker & Sijtsma, 2001) is the most restrictive CDM which involves only two item parameters regardless of the number of attributes required to correctly respond to a given test item. The main assumption is that the probability of success increases if an individual has mastered all the required attributes; otherwise, the absence of a required attribute cannot be made up for by the mastery of the other attributes. This model classifies individuals into two classes (2^k) for each item: (1) the first class consists of individuals who have mastered all of the attributes required to give a correct answer, and (2) the second class consists of individuals who have not mastered at least one of the required attributes. This assumption is known as conjunctive assumption. Furthermore, for each item, there are two item parameters: guessing (*g*) and slipping (*s*). If all the main effects and the lower order interaction effects are set to zero, the DINA model is obtained:

$$P(\alpha_{lj}^*) = \delta_{j0} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \qquad (2)$$

In this equation, $\delta_{j0} = g_j$ and $\delta_{j0} + \delta_{j12\ldots K_j^*} = 1 - s_j$.

2.2.3 DINO

DINO (Templin & Henson, 2006) is the compensatory counterpart of the DINA model. Similar to the DINA, there are two parameters for each item in the DINO model, and individuals are classified into two classes: (1) the first class includes individuals who have not mastered any of the required attributes measured by the item, and (2) the second group consists of individuals who have mastered at least one of the required attributes. The DINO model can be derived from the G-DINA model by constraining the magnitude of the main and interaction effects to be identical to each other and alternating the signs of the parameters which varies according to the order of interactions:

$$\delta_{jk} = -\delta_{jk'k''} = \cdots = (-1)^{K_{j+1}^*} \delta_{j12\ldots K_j^*} \qquad (3)$$

for k = 1, …, $K_j^* - 1$, and $k'' > k'$, …, K_j^*.

2.2.4 *A*-CDM

A-CDM (de la Torre, 2011) is an additive and a compensatory model. By setting all the interaction effects in the G-DINA model to zero, this model is derived. The *A*-CDM has an additive impact which is in contrast to the G-DINA which has a multiplicative impact. The main assumption in this model is that the mastery of each attribute additively and independently augments the probability of success, and deficiency in one attribute can be made up for by the mastery of other attributes. The *A*-CDM has $K_j^* + 1$ parameters for item j (de la Torre, 2011). The item response function (IRF) for the *A*-CDM is as follows:

$$P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} \quad (4)$$

2.2.5 LLM

As with the *A*-CDM, the LLM (Maris, 1999) is an additive and a compensatory model. When all the interaction effects in the G-DINA model are set to zero, LLM is obtained. In contrast to the *A*-CDM, a logit link function is used to estimate the LLM (de la Torre, 2011). The IRF for the LLM can be formulated as:

$$Logit[P(\alpha_{lj}^*)] = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} \quad (5)$$

which is the log-odds CDM without the interaction terms. It must be noted that the LLM is equivalent to the compensatory RUM.

2.2.6 RRUM

Another additive CDM is RRUM or Fusion Model (Hartz, 2002). Unlike the *A*-CDM, a log link function is used to estimate the RRUM (de la Torre, 2011). However, similar to the *A*-CDM and LLM, there are $K_j^* + 1$ parameters for item j. The IRF of the model can be written as:

$$Log[P(\alpha_{lj}^*)] = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} \quad (6)$$

3. Method

3.1. Participants

Item response data was collected from 1152 candidates for the Islamic Azad University English Proficiency Test (IAUEPT), an Iranian necessary language proficiency test, which was administered in April 2019. All candidates were Ph.D. students in different fields of study who intended to continue their studies in the Islamic Azad University (IAU) as a requirement for them to successfully graduate from the Ph. D program. The IAU Testing Centre I Teheran designs, develops, and administers this proficiency test at the same time in many cities across the country. The procedure to administer the test is carefully and meticulously designed and controlled by the representatives of the center who are present in every examination center in each city. Unfortunately, information about sthe age, gender, field of study, etc. of the participants is not available.

3.2. Instrumentation

The test analyzed in this study is the reading comprehension section of the IAUEPT. The test includes three sections of vocabulary (25 items), grammar (25 items), finding incorrect sentences (15 items), reading comprehension (20 items), and cloze (15 items). Test takers should answer all the questions (four-option multiple-choice) within 140 minutes.

The reading section includes two reading texts. Each of them consisted of 10 items. The first text had a 486-word text on Viola Desmond, who was an African Canadian woman from Nova Scotia. The readability score of the passage was 84.8 on Flesch Reading Ease Score scale and 6.5 on Gunning Fox scale; it was considered to be an easy text to read. The second text, about 196 words, was on migration. Its readability score was 51.8 on Flesch Reading Ease Score scale and 12.9 on Gunning Fox scale; it was considered by the readability formulas to be fairly hard to read. Cronbach alpha coefficient of the test was estimated, and the value was 0.63, indicating a moderate internal consistency.

3.3. Q-matrix

The Q-matrix used in the current study was taken from Boori et al. (2023). Q-matrix is a critical element of CDMs which specifies what attributes are required for each test item (Tatsuoka, 1983). In a recent study, they constructed and validated a Q-matrix for the reading comprehension section of the IAUEPT. They took the following steps to develop a Q-matrix. First, an initial Q-matrix was developed on the basis of L2 reading comprehension theories, previous studies on the application of CDMs to L2 reading comprehension, and brainstorming of four experienced university instructors (referred to as expert judges) with at least ten years of experience to determine the association between each item and its required attributes. Second, the initial Q-matrix was empirically validated using the strategy recommended by de la Torre and Chiu (2016). The modifications for the initial Q-matrix suggested by the software were carefully evaluated by the instructors. The suggested modifications were applied in the Q-matrix if they accorded to theories of reading comprehension and test domains; otherwise, they were disregarded. Boori et al. (2023) also checked the mesa plots for the suggestions as well as the Heatmap plot for identifying dependency between item pairs. Furthermore, the fit of the G-DINA model based on the initial Q-matrix was checked. After applying theoretically sound suggestions, a final Q-matrix was developed. Boori et al. (2023) explored five attributes underlying the reading comprehension section of the IAUEPT: vocabulary (VOC), grammar (GRM), making an inference (INF), understanding specific information (USI), and identifying explicit information (IEI). Table 1 demonstrates the final Q-matrix. For a detailed process of Q-matrix construction and validation, refer to Boori et al. (2023).

Table 1.
Final Q-matrix

Items	VOC	GRM	INF	USI	IEI
1	1	0	0	1	1
2	1	0	0	1	1
3	1	0	0	1	1
4	1	0	0	0	1
5	1	0	0	1	0
6	1	0	1	0	0
7	1	1	0	0	1
8	1	1	0	0	1
9	1	1	0	0	1
10	1	1	0	0	1
11	1	1	0	0	1
12	1	1	0	1	0
13	1	1	0	0	1
14	1	1	0	0	1
15	1	0	1	0	0
16	1	1	0	1	1
17	1	1	0	0	1
18	1	1	0	0	1
19	1	1	0	0	1
20	1	1	0	0	1

4. Data Analysis

The GDINA package version 2.8.8 (Ma et al., 2022) in R (R Core Team, 2022) was used to analyze the data. For this study, a two-stage analysis method was adopted. The GDINA package can generate a set of fit indices which can be employed to examine the extent to which the model fits the data (absolute fit statistics) or compare several models with each other to choose the best fitting model

(relative fit statistics). In the first stage, several absolute fit statistics were employed to compare the fit of the constrained models to the data against the G-DINA model at test-level. The following absolute fit indices were evaluated: (1) M2 (Chen & Thissen, 1997) is an averaged difference between the model-predicted and observed item response frequencies. A p-value higher than 0.05 indicates that the test items are independent, and the model has adequate fit to the data (Hu et al., 2016); (2) RMSEA2 (the root mean square error of approximation fit index for M2) is a measure of discrepancy between the observed covariance matrix and model-implied covariance matrix for each degree of freedom (Chen, 2007, p. 467). Values below 0.05 show satisfactory fit (Maydeu-Olivares & Joe, 2014). According to Hooper et al. (2008), models with RMSEA2 values less than 0.06 are models with sufficient fit; (3) The standardized root mean squared residual (SRMSR) is the square root of the sum of the squared differences of the observed correlation and the model-expected correlation of all item pairs (Chen, 2007). As argued by Maydeu-Olivares (2013, p. 84), values less than 0.05 indicate insignificant amount of misfit. Hu and Bentler (1999), however, showed that an ideal range for SRMSR is between 0 and 0.08. Overall, since there are no clear-cut criteria for most of the absolute fit statistics, researchers have argued that comparing the values of absolute fit statistics across a variety of CDMs can provide valuable information (Kunina-Habenicht et al., 2012; Lei & Li, 2016).

The fit of the G-DINA, as a general model, was also compared against a set of constrained CDMs like the DINO, DINA, LLM, A-CDM, and RRUM. The models were evaluated in terms of several relative fit indices: (1) log-likelihood is a function of sample size used to compare the fit of different coefficients (β). A higher value is better due to the presence of a tendency for maximizing the log- likelihood; (2) Akaike Information Criterion (AIC; Akaike, 1974) is used to select the best fitting model between non-nested models; and (3) Bayesian Information Criteria (BIC; Schwarz, 1978) is used to choose the best model between non-nested models. Models which have smaller information criteria are selected.

In the second stage, after exploring the most optimal model, the test takers' attribute mastery profiles, classification accuracies, and tetrachoric correlations between L2 reading attributes were examined. Finally, general and specific models were compared at item-level to evaluate whether, using the Wald test, the constrained models can replace the general model, e.g., the G-DINA model, without substantially losing model-data fit for each item. A main feature of general CDMs is that they permit each item to select the best model that best fits it. According to Ma et al. (2016), the Wald test statistic is calculated for each reduced model and then, when the p value is lower than $0 < 0.05$), the reduced model is not supported, and the G-DINA model is picked. However, when multiple reduced models are supported, and the DINA or DINO models are one of the retained models, the DINA or DINO model with the larger p-value is chosen as the best model. Nevertheless, if the DINA or DINO are not retained, the constrained model with the largest p-value is selected as the best model for this item. Bear in mind that when the p-values of several constrained models are greater than 0.05, the DINA or DINO models are preferred over the reduced models (e.g., A-CDM, LLM, and RRUM) due to their simplicity.

5. Results

5.1 Relative and Absolute Fit Statistics

The relative and absolute fit indices of the five constrained models against the G-DINA model are summarized in Table 2. Column two shows that the G-DINA model estimated 183 item parameters, DINA and DINO 71 parameters, and the A-CDM, LLM, and RRUM 108 parameters, indicating that the DINA and DINO are parsimonious models, and the G-DINA is the most complex model. In relation to the loglike and AIC, the lowest values were for the G-DINA, followed by the LLM, RRUM, *A*-CDM, DINA, and DINO. Because the G-DINA is a saturated model, its better fit was expected because the AIC always favors the saturated and more complex model (Li et al., 2015). With regard to the BIC, the LLM had the lowest value, and the RRUM was the closest model to it, succeeded by the *A*-CDM, G-DINA, DINA, and DINO. As argued by Li et al. (2015), the BIC prescribes a large penalty for more highly parameterized models, and this is the main reason why the value of the G-DINA model was high. Concerning M2 and SRMSR, the G-DINA had the best performance among the competing models, succeeded by the LLM, RRUM, *A*-CDM, DINO, and DINA. Also, the G-DINA had the lowest value in terms of SRMSR, followed by the RRUM and LLM. The DINA and DINO were the worst models, respectively. Finally, the last three rows show the results of the likelihood ratio test (LRT). The *p*-values

showed that none of the reduced models had equal fit as the G-DINA model, indicating that the G-DINA can better describe the structure of the test. The G-DINA model was thus selected for further analyses.

Table 2.
Relative and Absolute Fit Indices

	G-DINA	DINA	DINO	*A*-CDM	LLM	RRUM
NPAR	183	71	71	108	108	108
logLik	-13347.41	-13803.10	-13808.16	-13578.85	-13505.77	-13539.83
AIC	27060.81	27748.20	27758.33	27373.71	27227.54	27295.66
BIC	27984.83	28106.70	28116.82	27919.03	27772.86	27840.98
M2 (*p*)	31.7 (0.245)	526 (0)	517 (0)	253 (0)	197 (0)	198 (0)
SRMSR	0.0308	0.0559	0.0551	0.0403	0.0367	0.0364
RMSEA2	0.0122	0.0492	0.0486	0.0359	0.0284	0.0285
RMSEA2. CI1	0	0.0448	0.0441	0.0304	0.0224	0.0225
RMSEA2. CI2	0.0271	0.0537	0.0531	0.0415	0.0343	0.0344
Chisq	-	911.38	921.51	462.89	316.73	384.84
df	-	112	112	75	75	75
***p*-value**	-	<0.001	<0.001	<0.001	<0.001	<0.001

Note. NPAR = Number of parameters

5.2 Attribute Accuracy

Table 3 provides the classification accuracy (CA) at test- and attribute-level. CA refers to "the degree to which the classification of student latent classes based on the observed item response patterns agrees with students' true latent classes" (Cui et al., 2012, p. 23). In fact, it indicates to what extent examinees are precisely grouped into their true latent classes. As given in Table 3, the value of test-level accuracy is 0.77, suggesting a satisfactory accuracy of the test. The values of CA at attribute-level for the G-DINA were also higher than 0.85. This indicates a high degree of accuracy in classifying examinees into different (Cui et al., 2012; Wang et al., 2015).

Table 3.
Test- and Attribute-Level Accuracy

	Attribute-level Accuracy					Test-level Accuracy
G-DINA	**VOC**	**GRM**	**INF**	**USI**	**IEI**	
	0.95	0.89	0.96	0.90	0.87	0.77

5.3 Attribute Mastery Profiles

CDMs generally classify test takers into 2^k latent classes on the basis of the total number of attributes. For the present study, regarding the number of attributes, there exist 32 latent classes (2^5 = 32). In Table 4, 1s indicate that examinees have mastered the requisite attributes, and 0s show that

examinees have not mastered the necessary attributes. As an illustration, the profile [01011] demonstrates that the examinee has mastered the second, fourth, and fifth attributes (e.g., GRM, USI, IEI), and the first and third attributes (e.g., VOC and INF) have not been mastered by the examinee. As Table 4 presents, a large proportion of examinees have been classified into latent classes 25, 12, and 10 with 46%, 8%, and 8% attribute probabilities, respectively.

Table 4.
Proportion of Attribute Profile Patterns

Latent Class	Attribute Profile	G-DINA	Latent Class	Attribute Profile	G-DINA
1	00000	0.029	**17**	11100	0.000
2	10000	0.000	**18**	11010	0.069
3	01000	0.016	**19**	11001	0.004
4	00100	0.008	**20**	10110	0.000
5	00010	0.054	**21**	10101	0.000
6	00001	0.022	**22**	10011	0.061
7	11000	0.000	**23**	01110	0.006
8	10100	0.040	**24**	01101	0.009
9	10010	0.000	**25**	01011	0.460
10	10001	0.083	**26**	00111	0.000
11	01100	0.008	**27**	11110	0.000
12	01010	0.084	**28**	11101	0.000
13	01001	0.000	**29**	11011	0.024
14	00110	0.011	**30**	10111	0.000
15	00101	0.000	**31**	01111	0.000
16	00011	0.000	**32**	11111	0.010

5.4 Attribute Prevalence

Table 5 gives the difficulty of the four attributes. As it can be seen, making an inference (INF) and vocabulary (VOC) were the most difficult reading attributes for the test takers of the IAUEPT exam because about 9% and 29% of the test takers have not mastered these attributes, respectively. However, understanding specific information (USI) has been mastered by about 77% of the test takers which suggest that it is the easiest attribute, followed by the grammar (GRM) and identifying explicit information (IEI) with 69% and 67% attribute probabilities, respectively.

Table 5.
Attribute Prevalence

Attributes	Attribute Probability 1
Vocabulary	0.293
Grammar	0.690
Making an Inference	0.095
Understanding Specific Information	0.778
Identifying Explicit Information	0.671

5.5 Tetrachoric Correlations of L2 Reading Attributes

Table 6 shows the tetrachoric correlations among L2 reading attributes entailed in the reading section of the IAUEPT test. As can be seen, there are mostly negative correlations between the attributes, ranging from weak to moderate. There is a moderate negative correlation coefficient between INF and USI (-0.67), followed by INF and IEI (-0.62). On the contrary, there exists a strong positive correlation

between GRM and USI (0.84). This can be considered that if an examinee has a good performance on GRM, he/she has a higher probability to have a good performance on USI as well and vice versa.

Table 6.
Tetrachoric Correlations between L2 Reading Attributes

	VOC	GRM	INF	USI	IEI
VOC	1.00				
GRM	-0.66	1.00			
INF	0.36	-0.46	1.00		
USI	-0.53	0.84	-0.67	1.00	
IEI	-0.11	0.32	-0.62	0.26	1.00

5.6 Item-level Model Fit

The results of item-level model selection is demonstrated in Table 7. It can be seen that because all items of the test are multi-attribute (e.g., each item entails more than one attribute), some items variously selected reduced CDMs. In other words, for some items, the G-DINA model can be replaced by reduced CDMs without considerable loss in model data fit for each item. As can be seen, nine items picked the RRUM (e.g., Items 1, 7, 8, 9, 13, 14, 15, 16, and 19), eight items picked the G-DINA (e.g., Items 2, 5, 10, 11, 12, 17, 18, and 20), and three items picked the LLM (e.g., Items 3, 4, and 6).

Table 7.
Model Selection at Item-level

Items	Attributes	DINA	DINO	*A*-CDM	LLM	RRUM	Selected Model
1	VOC-USI-IEI	0.000	0.000	0.000	0.505	1.000	RRUM
2	VOC-USI-IEI	0.000	0.000	0.000	0.000	0.000	G-DINA
3	VOC-USI-IEI	0.000	0.000	0.005	0.999	0.961	LLM
4	VOC-IEI	0.000	0.000	0.005	0.434	0.003	LLM
5	V0C-USI	0.000	0.000	0.003	0.035	0.027	G-DINA
6	VOC-INF	0.000	0.000	0.674	0.996	0.091	LLM
7	VOC-GRM-IEI	0.000	0.000	0.000	0.720	1.000	RRUM
8	VOC-GRM-IEI	0.000	0.000	0.000	1.000	1.000	RRUM
9	VOC-GRM-IEI	0.000	0.000	0.052	1.000	1.000	RRUM
10	VOC-GRM-IEI	0.000	0.000	0.000	0.000	0.001	G-DINA
11	VOC-GRM-IEI	0.000	0.000	0.000	0.000	0.000	G-DINA
12	VOC-GRM-USI	0.007	0.000	0.000	0.010	0.033	G-DINA
13	VOC-GRM-IEI	0.000	0.000	0.000	0.163	0.641	RRUM
14	VOC-GRM-IEI	0.000	0.000	0.000	1.000	1.000	RRUM
15	VOC-INF	0.000	0.000	0.102	0.116	0.980	RRUM
16	VOC-GRM-USI-IEI	0.000	0.000	0.000	0.000	0.413	RRUM
17	VOC-GRM-IEI	0.000	0.000	0.002	0.012	0.014	G-DINA
18	VOC-GRM-IEI	0.000	0.000	0.000	0.000	0.003	G-DINA

19	VOC-GRM-IEI	0.000	0.000	0.000	0.952	1.000	RRUM
20	VOC-GRM-IEI	0.000	0.000	0.000	0.003	0.000	G-DINA

6. Discussion

The purpose of the present study was to firstly identify the most optimal CDM which can better explain the way attributes that underlie the reading comprehension section of the IAUEPT interact to yield correct responses, and then identify strong and weak points of the reading proficiency of the test takers. For the first purpose, the G-DINA model was compared against the other five models mentioned above. The model comparison was conducted with regard to absolute and relative fit indices. The G-DINA had the best performance on the basis of almost all the statistics, followed by the LLM, RRUM, *A*-CDM, DINA, and DINO. The *p*-values for LRT revealed that none of the reduced models had sufficient fit to the data compared to the G-DINA model. This indicates that, for the entire test, the G-DINA was the preferred model, so it was selected for further analyses. The analysis of the accuracy of the classification at the two levels of the attributes and the test confirmed the adequate fit of the G-DINA model. The results showed high and acceptable values for both levels, suggesting the precise classification of the examinees.

As the analysis of the attribute mastery profiles showed, the reading section of the IAUEPT exam has adequate diagnostic power to differentiate between those examinees who have mastered the necessary attributes and those who do not across all the items because only a low proportion of examinees have been put into the flat profiles. This is in disagreement with earlier research on L2 reading (e.g., Chen & Chen, 2016; Ravand, 2015) which reported that the two "flat" attribute mastery profiles (e.g., "non-master of all attributes" and "master of all attributes") were the most common classes. Lee and Sawaki (2009b) argued that when there exist high positive correlations among attributes, and the test is unidimensional, flat profiles appear. The reading comprehension part of the test could partition test takers into various classes in the current research.

Moreover, the results of the study showed that making an inference (INF) and vocabulary (VOC) are the hardest attributes, and understanding specific information (USI) is the easiest one, followed by grammar (GRM) and identifying explicit information (IEI). Some researchers have argued the presence of a hierarchical association among reading attributes (Baghaei & Ravand, 2015; Ravand, 2015). Harding et al. (2015) state that making an inference, understanding main idea, identifying explicit information, and understanding detailed information are higher-level attributes, and vocabulary and grammar are the lower- level reading comprehension attributes. For this reason, since INF requires understanding both explicit and implicit meanings of a given text and thus involves higher level processing of information, it can be assumed to be a hard attribute for the test takers (Grabe, 2009). However, the second most difficult attribute was VOC which is not in accordance with earlier studies. This could be due to the structure of the reading comprehension section of the IAUEPT exam because most of the items require vocabulary knowledge. Another reason might be the fact that most IAUEPT test takers have very poor knowledge and ability in English, especially vocabulary. They are in fact the beginners who are expected to pass an upper-intermediate test without any systematic and pedagogically adequate training. In fact, test takers are expected to have a large repertoire of lexical knowledge to be able to respond correctly to the given items. Furthermore, another finding of the present study which diverges with previous studies is that the USI as a higher level reading attribute was the second easiest attribute. The easiness of the USI can be attributed to the co-existence of this attribute and IEI in most of the IAUEPT test items and format of the items which tap the ability of test takers to find specific information. In other words, test items which require the presence of the USI are not powerful enough to make a distinction between the two attributes.

In addition, the results of tetrachoric correlations showed that there are mainly weak to moderate negative correlation coefficients between reading attributes. This can be considered as evidence that a test taker requires to have a mastery of all the attributes tapped by an item to get the item right. On the other hand, there was a strong positive correlation between GRM and USI. This indicates that if a test taker has a good performance on GRM, he/she has a higher probability to have a good performance on USI as well and vice versa.

Finally, the results of model selection at item-level indicated that for all items of the test which require more than one attribute, three items picked the LLM (e.g., Items 3, 4, and 6), eight items selected the G-DINA (e.g., Items 2, 5, 10, 11, 12, 17, 18, and 20), and nine items selected the RRUM (e.g., Items 1, 7, 8, 9, 13, 14, 15, 16, and 19). This finding agrees with earlier research on L2 reading which suggested a combination of interactions among L2 reading attributes (Li et al., 2015). An important piece of information for the analysis at item-level is that some items that tap multiple attributes selected the G-DINA as the best fitting model.

Another noteworthy point is that none of the items picked the *A*-CDM, DINA, and DINO models. This tells us that the modeling structure of these three models do not fit the reading section of the IAUEPT test, of course with this Q-matrix and dataset. In the *A*-CDM, it is assumed that the mastery of each attribute additively augments the probability of success, and defect in an attribute can be compensated for by the mastery of other attributes.

Furthermore, the DINA and DINO models are too restrictive for explaining reading comprehension attributes (Yi, 2017; Li et al., 2015). DINA is considered as an overly restrictive and non-compensatory model in the sense that a test taker must possess all the attributes that are required to give a correct answer. The examinee cannot succeed if he/she has possessed only one of the required attributes. On the other hand, the DINO is considered as the most extreme case of compensation in the sense that an examinee who possesses only one of the required attributes has the same probability of success as examinees who possess all the required attributes. In the DINO model, the presence of only an attribute can make up for the non-mastery of the other attributes. The results in the present study suggest that these two parsimonious models are too simple to describe the structure of the reading section of the IAUEPT exam, that is, L2 reading comprehension cannot be limited to two extreme cases of compensatory and non-compensatory manner.

7. Conclusion

The findings of this study have numerous implications in terms of both theory and practice for all stakeholders. Theoretically, the findings will extend the literature on retrofitting existing non-diagnostic tests, especially on Iranian high-stakes tests, with the application of CDMs, and diagnosing candidates of IAUEPT exam. Understanding the nature of attributes that underlie L2 reading comprehension would also allow researchers to develop more logical and robust reading comprehension theories and models (Buck & Tatsuoka, 1998).

Pedagogically, all educational scholars and stakeholders can benefit from the results of the present study. For example, test developers can use the results of this study to construct items which can provide better and more detailed diagnostic information about the performance of test takers on the IAUEPT test. The diagnostic information provided by this study can help teachers in preparation courses and colleges to pay more attention to attribute mastery probabilities for individual candidates and overall groups. Teachers can use such information to tailor or improve their lesson plans to satisfy the needs of candidates with various proficiency levels, provide effective educational materials or enhance the quality of them so as to help individual candidates to reduce and remove their weaknesses, and ultimately improve the process of teaching and learning. More importantly, students themselves can take advantage of the current study's results. Giving feedback to learners is the major purpose of diagnostic testing (Harding et al., 2015). Diagnostic feedback enables students to understand the strong and weak points of the reading proficiency of the test takers in different attributes of L2 reading comprehension to adopt some strategies for the purpose of improving their reading comprehension ability. As Black and Wiliam (1998) highlighted, diagnostic feedback should be descriptive and interpretable to allow examinees to make the gap between their current proficiency level and their desired level smaller.

Like every other research, this research also encountered some limitations. One limitation is that we utilized a non-diagnostic test in this research to elicit diagnostic information about the performance of test takers on a large-scale test. As Jang (2009) noted, this retrofitting analysis of existing tests is problematic because the inferences made based on test takers' attribute mastery profiles would be unreliable. However, Lee and Sawaki (2009a) argued that retrofitting existing non-diagnostic tests could save time and budget for developing a cognitive diagnostic test. It is highly recommended for future studies to design true diagnostic tests based on a CDA framework.

The second limitation is that the researcher did not have access to the subjects' demographic information such as age, gender, educational background, field of study, and language learning profile. The demographic information could have allowed researchers to examine differential item functioning (DIF). Having access to the results of all the subjects taking the same version of the test or other versions of the test and comparing the acquired diagnostic information from different administrations would offer more reliable and beneficial results.

One area for further investigation is how characteristics of examinees, such as gender, learning style, proficiency level, etc., and grain size of attributes can impact the efficiency of diagnostic feedback. In a similar vein, it is recommended for studies focusing on model comparison to include both low and high levels of proficiency in order to examine whether the inter-attribute relationships differ across various proficiency levels (Ravand & Robitzsch, 2018). As stated by Alderson (2000), reader and text variables are the main two factors which affect reading comprehension. When the reading passage is used for assessment purposes, test-related factors as another important source that affects reading comprehension is also involved (Ravand & Robitzsch, 2018, p. 20).

The final recommendation is that future studies can empirically investigate the effect of providing diagnostic information to all stakeholders. Due to logistical problems and the use of a borrowed dataset, this study was not able to give a diagnostic report to all test takers. Future studies can examine to what extent diagnostic feedback could help stakeholders. According to Cumming (2015, p. 414), "instructors and administrators charged with preparatory or supplementary courses or other activities are the people who truly need detailed, relevant, and valid information from diagnostic language assessments to design, implement, evaluate and refine their courses or other activities." Providing stakeholders with diagnostic information would allow them to adopt some strategies to remedy weaknesses.

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.

Funding (obligatory)

The author(s) received no specific funding for this work from any funding agencies.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716−723. http://dx.doi.org/10.1109/TAC.1974.1100705

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in in English as a foreign language using the 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

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice, 5*(1), 7−74. https://doi.org/10.1080/0969595980050102

Boori, A. A., Ghazanfari, M., Ghonsooly, B., & Baghaei, P. (2023). The construction and validation of a Q-matrix for cognitive diagnostic analysis: The case of the reading comprehension section of the IAUEPT. *International Journal of Language Testing, Special Issue*, 31−53. https://doi.org/10.22034/ijlt.2023.383112.1227

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/10.1191/026553298667688289

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, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly, 13*(3), 218−230. https://doi.org/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. https://doi.org/10.2307/1165285
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
Cumming, A. (2015). Design in four diagnostic language assessments. *Language Testing, 32*(3), 407−416. https://doi.org/10.1177/0265532214559115
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
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
Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179−197. https://doi.org/10.1037/0033-2909.93.1.179
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
Gottardo, A., & Mueller, J. (2009). Are first- and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology, 101*(2), 330−344. http://dx.doi.org/10.1037/a0014320
Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *RASE: Remedial & Special Education, 7*(1), 6−10. https://doi.org/10.1177/074193258600700104
Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.
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. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x
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.
Helm, C., Warwas, J. & Schirmer, H. (2022). Cognitive diagnosis models of students' skill profiles as a basis for adaptive teaching: an example from introductory accounting classes. *Empirical Research in Vocational Education and Training, 14*(9), 1–30. https://doi.org/10.1186/s40461-022-00137-3
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.html
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
Hooper, D., Coughlan, J., & Mullen, M. (2008, June). *Evaluating model fit: a synthesis of the structural equation modelling literature*. In 7th European Conference on research methodology for business and management studies (pp. 195−200).
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.). Cambridge University Press.
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., & 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
Javidanmehr, Z., & Anani Sarab, M. R. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA model to a high stakes reading comprehension test. *Language Assessment Quarterly, 16*(3), 294−311. https://doi.org/10.1080/15434303.2019.1654479
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, A.-Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258. https://doi.org/10.1177/0265532214558457
Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 209−226). Blackwell Publishing.
Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press.
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/10.15434300903079562
Lee, Y. W., & Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172−189. https://doi.org/10.15434300902985108
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
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., & 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. https://doi.org/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*(3), 391−409. https://doi.org/10.1177/0265532215590848

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). *GDINA: The generalized DINA model framework*. R package version 2.8.8. Retrieved from https://CRAN.R-project.org/package=GDINA

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187−212. https://doi.org/10.1007/BF02294535

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71−101. https://doi.org/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. https://doi.org/10.1080/00273171.2014.911075

McNeil, L. (2012). Extending the compensatory model of second language reading. *System, 40*(1), 64−76. https://doi.org/10.1016/j.system.2012.01.011

Mehrazmay, R., Ghonsooly, B., & de la Torre, J. (2021). Detecting Differential Item Functionin Using Cognitive Diagnosis Models: Applications of the Wald test and likelihood ratio test in a university entrance examination. *Applied Measurement in Education, 34*(4), 262−284. https://doi.org/10.1080/08957347.2021.1987906

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed., pp. 1−103). American Council on Education/Macmillan.

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 University Press.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

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. https://doi.org/10.1016/j.stueduc.2017.10.007

Ravand, H. (2015). 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. (2019). 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

Sadoski, M., & Paivio, A. (2007). Toward a unified theory of reading. *Scientific Studies of Reading, 11*(4), 337−356. https://doi.org/10.1080/10888430701530714

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461−464. https://doi.org/10.1214/aos/1176344136

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*(1), 99−128. https://doi.org/10.1177/0265532207071513

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. (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 misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345−354. Retrieved from: https://www.jstor.org/stable/1434951

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

Usó-Juan, E. (2006). The compensatory nature of discipline-related knowledge and English language proficiency in reading English for academic purposes. *The Modern Language Journal, 90*(2), 210−227. https://doi.org/10.1111/j.1540-4781.2006.00393.x

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, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement, 48*(2), 165−187. https://doi.org/10.1111/j.1745-3984.2011.00142.x

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

Weir, C., Yang, H. Z., & Jin, Y. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes*. Cambridge University Press.

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. https://doi.org/10.1177/0265532216646141

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal, 96*(4), 558−575. https://doi.org/10.1111/j.1540-4781.2012.01398.x