Tabaran Institute of Higher Education ISSN 2476-5880
International Journal of Language Testing
Vol. 10, No. 1, March 2020

A Cognitive Diagnostic Modeling Analysis of the English Reading Comprehension Section of the Iranian National University Entrance Examination

Seyed Jamal Hemati[1] and Purya Baghaei[2]

Received: 10 April 2019 Accepted: 23 December 2019

Abstract

Cognitive Diagnostic Models are a class of multidimensional categorical latent trait models which provide diagnostic information by reporting examinees' mastery profiles on a set of predefined skills. CDMs provide fine grained information concerning examinees' strengths and weaknesses in the subskills and subprocesses which constitute a larger domain of knowledge. Such detailed information helps in classroom teaching, designing remedial courses, and material development. In this study, we analysed a high stakes English as a foreign language reading comprehension test using GDINA model. The skill profiles of the test takers, the class probabilities, attribute mastery probabilities, attribute difficulties, and model data fit at test and item level were examined. Implications of the study for reading comprehension research and CDM applications are discussed.
Keywords: Cognitive Diagnostic Model, reading comprehension test, National University Entrance Examination

1. Introduction

In Traditional large scale testing usually item response theory models are used to scale examinees and compare schools, districts, and countries. Such psychometric models provide individual ability parameters on a latent continuum which is appropriate for the purpose of ranking and estimating test takers' general abilities. However, modern assessment demands necessities more than a single generic score on a latent ability. Stakeholders require categorical classification of respondents, such as master/nonmaster, on certain abilities or attributes rather than a parameter on a continuous scale. Providing formative diagnostic feedback to stakeholders to inform them of students' weaknesses and strengths or diagnosing disorders which are indicated by the presence or absence of certain syndromes in patients and

[1] English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran. *(Corresponding author)* Email: Jamalhemmati@gmail.com

[2] English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran.

classifying them as 'healthy' or 'clinically ill' is deemed essential for developing targeted intervention (Rupp & Templin, 2008).

Cognitive Diagnostic Models are a class of multidimensional categorical latent trait models which provide fine-grained diagnostic information by reporting examinees' mastery profiles on a set of predefined skills (DiBello, Stout, & Roussos, 1995; Hartz, 2002). CDMs are the intersection between psychometrics and cognitive science which promote *assessment for learning* as opposed to the more traditional role of the *assessment of learning outcomes*. The major difference between CDMs and multidimensional IRT models and confirmatory factor analysis is that the latent trait in CDMs is conceptualized as categorical while in the latter models is continuous (Effatpanah, 2019; Ravand & Baghaei, 2020; Ravand, Barati, & Widhiarso, 2013).

Rather than locating individuals on a continuous ability scale CDMs assign mastery/nonmastery classifications to individual examinees. Examinees who have the same total scores do not necessarily have the same strengths and weaknesses and might have different mastery profiles. CDMs provide fine grained information concerning examinees' strengths and weaknesses in the subskills and subprocesses which constitute a larger domain of knowledge. Such detailed information helps in classroom teaching, designing remedial courses, and material development.

CDMs are categorized under three broad categories of compensatory, noncompensatory, and general. Each of these types is different in modeling the relationship between the probability of a correct reply to an item and the mastery of the subskills constituting the item. Compensatory models assume that mastery of one or some of the attributes required to answer an item can make up for nonmastery of other attributes. Noncompensatory models, on the other hand, specify that for correctly answering an item all the required attributes for the item should be mastered. General CDMs allow for both types of relationships within the same test. They allow for multiple CDMs for different items. That is, general CDMs allow the researcher to hypothesize varying relationships among the attributes across the items. Therefore, the choice of a CDM should be specified in advance and be guided by the theory of the construct under investigation. According to de la Torre and Lee (2013), employing general models is helpful in that "(a) CDMs need not to be specified *a priori*, and (b) multiple, statistically determined CDMs can be used within a single assessment" (p.370).

As noted above the advantage of general CDM models is that they allow for different CDMs for the items within a test. de la Torre and Lee (2013) introduced the Wald test to objectively choose the best-fitting model for each multi-attribute item. The method developed by de la Torre and Lee evaluates the fit of the G-DINA, at the item level, against the fit of the DINA, DINO, and ACDM. The assumption of general models is more realistic as it is hard to postulate a similar relationship among the attributes across all items. The relationship between the attributes "might change depending on the difficulty of the attributes, the area of language tapped by the items, the cognitive load of the attributes…" (Ravand, 2016, p.13).

Nevertheless, von Davier (2014) using a simulation study demonstrated that the DINA model is equivalent to a general compensatory family of diagnostic models. Close fit and very high agreements between the parameter estimates from the DINA and the compensatory GDM, even when the data were simulated using DINA as the generating model, indicated that substantive assumptions about the nature of the relationships among the skills, i.e., whether compensatory or noncompensatory, as a result of fit of a certain model is not warranted.

> …the example data were fitted in identical ways by the DINA and equivalent DINA, so there is no way to decide which model generated the data. Given the existence of at least two linear (compensatory) DINA equivalent GDMs, do we really have evidence of the skills needed to solve the items that are conjunctive? (von Davier, 2014, p.68).

Several different types of CDM models have been advanced in the past few years. These models differ in the way they postulate the relationship between attributes and the probability of a correct response. de la Torre (2011) introduced the *generalized DINA* (G-DINA) model which is the general extension of the disjunctive DINA model. Like other CDMs, G-DINA requires a J×K Q-matrix. G-DINA partitions the examinees into $2^{K_j^*}$ latent classes where K_j^* is the number of required attributes denoted as α_{lk} for item j. Therefore, if there are four subskills or attributes covered in the test there will be (2^4) 16 latent classes. Assuming conditional independence as well as independence among the subjects, the G-DINA model is formally expressed as follows:

$$P\left(\boldsymbol{\alpha}_{lj}^*\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk},$$

where $P\ (\alpha_{lj}^*)$ is the probability that an examinee in latent class l gets item j, with attribute pattern α_{lj}^* , correct. If item j requires say, subskills 3 and 5 to get solved its attribute vector will be $\alpha_{lj}^*= (\alpha_{l3}, \alpha_{l5})$. δ_{j0} is the intercept of item j, i.e., the probability of a correct answer to an item when none of the required attributes for the item has been mastered. δ_{jk} is the main effect due to α_k; it is the change in the probability of a correct response as a result of mastering the attribute α_k. $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$; it is the probability of a correct response due to the mastery of both attributes that is above and beyond the influence of the simple additive impact of the two attributes. $\delta_{j12\ldots K_j^*}$ is the interaction effect due to α_1,…, $\alpha_{K_j^*}$; "It represents the change in the probability of a correct response due to the mastery of all attributes that is over and above the impact of the main and lower-order interaction effects" (García, Olea, & de la Torre, 2014, p. 373).

The model gives a vector of estimates of the expected a posteriori (EAP) probabilities of mastery of each subskill for each individual examinee. By convention, if the probability of a mastery of a subskill is greater than 0.50 the person is designated as a master of the subskill

and if the probability is smaller than 0.50 the person is a nonmaster of the subskill (de la Torre, Hong, & Deng, 2010; Effatpanah, Baghaei, & Boori, 2019; Ravand, Baghaei, & Doebler, 2020; Templin & Henson, 2006). In our hypothetical example with four subskills if the probabilities of mastery of these subskills for a person are (.40, .69, .85, .31) then this person's mastery profile would be (0, 1, 1, 0). This means that this person has mastered the second and the third subskills but not the first and the fourth. The model also reports the probability of each latent class and attribute profile and the number of examinees falling in each class. Furthermore, the probability of each examinee belonging to each latent class is also provided by the model. Computing the percentage of examinees who have mastered each attribute gives a measure of attribute difficulty.

The G-DINA parameters estimated by the model are: intercept, main effects, and interaction effects. The intercept parameter shows the probability of answering each item correctly when none of the attributes required by the item has been mastered. The main effects show the increase in the probability of correctly answering each item when only one of the attributes has been mastered, and the interaction effect shows the increase in the probability when a combination of the attributes has been mastered. For example, for items which require two attributes to get solved, four parameters are estimated: one intercept, two main effects for the attributes and one interaction effect. For items which require three attributes, eight parameters are estimated: one intercept, three main effects, and four interaction effects.

2. Literature Review

In this section previous applications of CDM's to second language (L2) reading comprehension tests are summarized and reported. In these studies CDMs have been applied to high stake tests such as the TOFEL, IELTS or large scale national university entrance examinations. It is also important to note that these tests are not developed from the beginning to extract diagnostic information about test takers' language abilities; therefore CDM's are retrofitted to the existing, non-diagnostic tests. Although Jang (2005) argues that the use of non-diagnostic tests for diagnostic purposes might negatively affect the results of the study, Lee and Sawaki (2009), believe that such retrofitting efforts could serve as an important step in advancing second language assessment research. Many of the applications of the CDMs (including this study) are retrospective studies in which non-diagnostic tests are used to extract diagnostic information (Ravand & Robitzsch, 2015).

Kim (2015) in her research on cognitive diagnostic assessment of L2 reading ability used the reading test data from the reading section of a placement test. The test was initially developed to place incoming students into classes at different proficiency levels in an adult ESL program. To identify the attributes required for answering the test items, previous literature on communicative language ability model (Bachman & Palmer, 1996) was used. Based on experts' consensus 10 attributes were identified. In the next step to develop the Q-matrix the experts were asked to indicate the major attributes required for responding to each item. The initial Q-matrix which focused on the conjunctive interaction of attributes was further refined through Fusion model (Hartz, Roussos, & Stout, 2002) in an iterative process

where the refined Q-matrix consisted of 10 L2 reading attributes. Each item measured between one to four attributes.

The ten attributes were: (1) Lexical meaning; (2) Cohesive meaning; (3) Sentence meaning; (4) Paragraph/text meaning; (5) Pragmatic meaning; (6) Identifying word meaning; (7) Finding information; (8) Skimming; (9) Summarizing, and (10) Inference making. As a result of her study she found out that knowledge of cohesive meaning was the most difficult and the knowledge of pragmatic meaning was the easiest attribute for participants. This finding was considered as evidence for hierarchy of L2 reading attributes. She also showed that the model had the power to categorize participants into three different levels of proficiency based on their mastery profiles: beginner, intermediate and advanced levels. In addition, examinees' strengths and weaknesses were identified for the overall group, three reading proficiency groups and individual learners.

In another study, Lee and Sawaki (2009) compared the results of three different cognitive diagnostic models when applied to reading and listening sections of two field test forms of TOFEL iBT. In this study, in addition to providing examinees' mastery profiles, they wanted to check the consistency of mastery classifications over different models as well as consistency of the results over two different forms of the exam. Three CDMs were the general diagnostic model (von Davier, 2005), the fusion model (Hartz et al., 2002), and the latent class analysis (Yamamoto, 1990). In latent class analysis and fusion model, non-compensatory interaction is assumed among attributes while in thegeneral diagnostic model such limitation is not imposed. In this study the Q-matrix was developed based on earlier research on content analysis of individual test items of the TOFEL iBT reading section. Four reading attributes/skills were identified including: (1) Understanding Word Meaning; (2) Understanding Specific Information; (3) Connecting Information; and (4) Synthesizing and Organizing Information. The results of this study indicated that: (1) all three models had the power to separate respondents into two mastery levels on most of the reading and listening skills; (2) Based on the list of reading and listening attributes defined in this study, a moderate level of across-form consistency of examinee skill mastery classification was achieved; (3) There were a considerable portion of examinees with "flat" profiles (master of all attributes or non-master of all attributes). Existence of a large portion of flat profiles was interpreted as evidence for unidimensionality of reading and listening sections, (4) Despite some minor differences, the three models led to almost the same results in terms of examinee mastery classification.

Ravand, Barati, and Widhia (2013) used the DINA model (de la Torre & Douglas, 2004) to mainly investigate the diagnostic capacity of a high stake reading comprehension test which was administered to PhD program applicants at the University of Isfahan, Iran. In their study to identify the attributes and development of the Q-matrix, they used experienced university instructors to brainstorm on the possible attributes measured by the test. Then two other university instructors were asked to independently specify the attributes required by each of the reading comprehension items. The final Q-matrix was developed based on experts consensus. The five identified attributes included: (1) vocabulary, (2) syntax, (3) extracting explicit information, (4) connecting and synthesizing, and (5) making inferences.

The results of this study indicated that a large portion of participants had flat mastery profiles. Diagnostically informative items are those with low slipping and guesting parameters and high item discrimination indexes (IDI) but the results of this research showed high slipping and guesting parameters and low IDI indexes which questioned the diagnostic value of Isfahan University reading comprehension test.

In another study, Ravand (2016) used the G-DINA model (de la Torre, 2011), to put more insight into attributes involved in answering reading comprehension items of the national University Entrance Examination (UEE) in Iran. The test was developed to select among graduate university students who desired to participate in English master programs at state-run universities in Iran. In this research the attributes were specified and the Q-matrix was developed based on experts' brainstorming on the possible subskills measured by the test and the attributes measured by each item. Then the initial Q-matrix was empirically validated and revised by the use of a discrimination index proposed by de la Torre and Chiu (2010). Ravand split the sample into two halves. One half was used to identify and revise the Q-matrix and the other half was used to cross-validate the obtained Q-matrix. Five attributes were specified in this study including: (1) reading for details, (2) reading for inference, (3) reading for main ideas, (4) syntax, and (5) vocabulary. The analysis provided detailed diagnostic information about participants' strengths as well as different latent mastery classes.

Another example of CDM use is the research done by Jang (2005) in which she first conducted a verbal protocol analysis to identify the L2 reading attributes. As the source of data for her study she used the reading section of TOFEL. Nine attributes were identified in this study including: (1) Deducing word meaning from the context, (2) Determining word meaning out of context, (3) Comprehending text through syntactic and semantic links, (4) Comprehension of text-explicit information, (5) Comprehending text-implicit information at global level, (6) Making inference about major arguments or a writer's purpose, (7) Comprehending negatively stated information, (8) Summarizing major ideas from minor details, and (9) Determining contrasting ideas through diagrammatic display. For the next step the Fusion model (Hartz, Roussos, & Stout, 2002) was used to specify attributes and participants' reading abilities.

Jang (2009) further used the results of her 2005 study and investigated the validity of the Fusion model for diagnostic analysis of the reading comprehension section of the LanguEdge assessment. The nine attributes which were identified in the previous research were used again as a basis for developing the Q-matrix. The results of her study supported the statistical quality of the Fusion model. The estimates from the model approximated the statistics observed from the real data which was evidence of a good fit. However a misfit of data at the two ends of distribution was considerable. She observed that through this model the scores of high-scorers were overestimated and the scores of low-scorers were underestimated; thus she concluded that mastery profiles for high and low scoring test takers may not be as accurate as those of the others and this type of mastery profiles might lead to inaccurate diagnostic feedback. She also claimed that the psychometric properties of the items of a test such as the TOFEL which is a norm-referenced test with the purpose of placing examinees on a

continuous scale may not be relevant to a diagnostic test. The results also indicated that the model-estimated probabilities of the mastery profiles were in line with participants' self-assessment about their own reading ability. Therefore she advocated the use of self-assessment for diagnostic purposes in combination with statistical diagnostic feedback.

L2 reading ability as one of the major language skills has been the subject of many studies and researchers try to shed light on different aspects of this complex language skill. Lots of research has been conducted on linguistic aspects of L2 reading ability but studies which address the cognitive aspects of this skill are rather new and still lots of efforts should be done in order to have a sound understanding of attributes involved in successful performance on L2 reading tests (Alderson, 2005).

The notion of subskills was introduced in the modern theories of language skills as the constituent elements which form a skill. The assumption is that by mastering the subskills and building a repertoire learners can effectively communicate (Goh & Aryadoust, 2014). By implication, deficiencies in the subprocesses which constitute a skill can lead to breakdowns in communication. Material development, teaching methodology, and the testing of foreign language proficiency is based on the subskills theory.

While measuring test takers' second language general proficiency or individual language skills is possible using traditional IRT modeling, accurate diagnosis of the subprocesses underlying each skill requires CDM. Diagnostic assessment of second language proficiency has gained lots of attention over the past decade (Alderson, 2005; Alderson, Brunfaut, & Harding, 2015; Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja, 2015). Application of CDM's to non-diagnostic tests by retrofitting to existing achievement and proficiency tests, provides a great wealth of information at the subprocess level that can be benefitted in the classroom and for designing remedial courses.

In the Iranian educational system high-stakes national tests are administered every year to admit candidates into state universities. The tests contain several components and examinees, depending on the progrmmme of their interest, take the appropriate tests. The English language test is a component which should be taken by all candidates regardless of the programme they want to peruse at university. The test is only used for estimating overall scores which shows the ranking of candidates in the national sample to make pass/fail decisions. Nevertheless, considering the size and the representativeness of the sample who take the test a great wealth of diagnostic information is contained in the data that is never investigated

The purpose of the present investigation is to take advantage of the valuable datasets available and analyse the reading comprehension section of the English test with the G-DINA model. Such an analysis will provide additional information to researchers and educators on the subskills and processes that are required to complete the reading test. The derived information can be used in classroom teaching, intervention programmes for improving reading comprehension in English as a foreign language, material development and syllabus design, and any decision for the amendment of the English curriculum.

3. Method

3.1 Participants and data source

The English language section of the Iranian University Entrance Examination (IUEE) for candidates who want to study foreign languages in state universities was used in this study. The data of 10000 participants who took the test in 2012 are analyzed. The test contains 70 four-option multiple-choice items and should be answered in 105 minutes. The test contains six sections: grammar (10 items), vocabulary (15 items), sentence structure (recognizing the correct structure of sentences; 5 items), language function (the ability to use English in real interactions; 10 items), cloze (10 items), and reading comprehension (20 items). In this study only the reading comprehension section is analyzed. The reading section contains three passages with varying lengths between 425 to 495 words.

3.2 Construction of the Q-matrix

To define the attributes involved in a test, different sources can be used, including theories of content domain, test specifications, content analysis of the test items, and think-aloud protocol analysis (Hemmati, Baghaei, Bemani, 2016; Leighton, Gierl, & Hunka, 2004; Leighton & Gierl, 2007). In this study we used the data of a non-diagnostic test to extract diagnostic information about test takers' comprehension reading ability (retrofitting case). There was neither test specifications nor detailed cognitive model of task performance available.

To determine the attributes that the candidates should have mastered in order to answer the reading comprehension test items, in the initial step researchers consulted the literature on language ability models in which the reading comprehension attributes and subskills are discussed. The models reviewed included the model proposed by Hughes (2003) consisting of 20 attributes including (1) Identify pronominal references, (2) Identify discourse markers, (3) Interpret complex sentences, (4) Interpret topic sentences, (5) Outline logical organization of a text, (6) Outline the development of an argument, (7) Distinguish general statements from examples, (8) Identify explicitly stated main ideas, (9) Identify implicitly stated main ideas, (10) Recognize writer's intention, (11) Recognize attitudes and emotions of the writer, (12) Identify addressee or audience for a text, (13) Identify what kind of text is involved (e.g. editorial, diary, etc.), (14) Distinguish fact from opinion, (15) Distinguish hypothesis from fact, (16) Distinguish fact from rumor or hearsay, (17) Infer the meaning of an unknown word from the context, (18) Make propositional informational inferences answering questions beginning with who, when, what, (19) Make propositional explanatory inferences concerned with motivation, cause, consequence and enablement, answering questions beginning with why and how, and (20) Make pragmatic inferences. This model is an instant of a comprehensive model in which all the probable reading attributes are mentioned.

Another model which was consulted in this study is proposed by Farhadi, Ja'farpour, and Birjandi (1994). This model consists of a narrower domain of attributes for reading comprehension. These attributes are: (1) Guess the meaning of words from context, (2) Understand the syntactic structure of the passage, (3) Get explicit and implicit ideas, (4)

Grasp the main idea of the passage, (5) Recognize the tone, mood and purpose of the writer, (6) Identify literary techniques of the writer, and (7) Draw inferences about the content of the passage.

Other useful sources for identifying the attributes are the previous research conducted in the field of reading comprehension. Therefore, we referred to numerous studies in which the reading comprehension attributes were investigated (Buck et al., 1997; Clark, 2013; Jang, 2005; Kim, 2014; Lee and Sawaki, 2009a; Ravand, 2016; Ravand & Robitzsch, 2018; Ravand, 2013; Sheehan, 1997; Svetina et al., 2011; Vander Veen et al., 2007;; Zheng & De Jong, 2011;). At this stage we had a pool of reading comprehension subskills and attributes among which we had to choose the attributes required for answering the test items. To this end, the authors brainstormed on the possible attributes measured by the test. From the pool of attributes those which seemed to be required for answering the test items based on the content analysis of each item were selected. A set of five attributes underlying the reading test was specified, including (1) Making inferences, (2) Extracting explicit information, (3) Identifying word meaning from the context, (5) Identifying pronominal references, and (5) Evaluating response options.

For the next step, two other English teachers with over five years of teaching reading comprehension experience were asked to independently specify the attributes measured by each of the 20 reading comprehension items. Finally in a session, the authors and the two English teachers came together and agreed on the attributes which were required for answering the reading comprehension items. For the next step, to develop the Q-matrix the experts were asked to assign the attributes to each of the items. To construct the Q-matrix the team of experts including the researchers and the two experienced teachers, separately, selected among the five attributes defined in the previous step and assigned them to the items. Then they checked the Q-matrices together and came to an agreement about the Q-matrix.

In the next step, the Q-matrix was subjected to statistical analysis through the procedure proposed by de la Torre and Chiu (2016) using the GDINA package (Ma et al., 2016) in R. The procedure is based on a discrimination index which measures the degree to which an item discriminates among different reduced q-vectors and can be used in conjuction with the G-DINA and all the constrained models subsumed under it. de la Torre and Chiu's procedure identifies potential misspecifications and provides suggestions for modification of the Q-matrix. The suggested modifictions are either turning 0 entries into 1s or vice versa.

According to the validation procedure 11 of the 0's in the final Q-matrix had to be converted to 1. To include human logic into the Q-matrix specification model suggestions were compared with the independent coding of the experts and only those which had also been suggested by at least one of the coders in their independent codings were implemented. Six of the 11 transformations suggested by the analysis had also been suggested by the human raters. Therefore, only these modifications were implemented only.

4. Analyses

Model Fit

The parameters obtained from CDM models are interpretable to the extent that the model fits the data. To check the model fit there are two common methods: 1) to check the *absolute fit*, in which we check the model to the data, and 2) to check the *relative fit*, which is comparing the results of a model to the other rival models (Ravand & Robitzsch, 2015). Through the comparison of observed and model-predicted response frequencies of item pairs, a range of absolute fit indices are obtained (Maydeu-Olivares, 2013). The absolute fit indices are illustrated in Table 1.

Table 1*:* Absolute Fit Indices

Model	MADcor	SRMSR	MADRESIDCOV (MADRCOV)	MADQ3
GDINA	0.05	0.06	0.56	0.04

The absolute fit indices, presented in Table 1 are defined as follows:

- MADcor: The mean absolute difference for the item-pair correlations (DiBello, Roussos, & Stout, 2007). It is the difference between the model-predicted and the observed item correlations.
- SRMSR: standardized root mean squared residuals.
- MADRESIDCOV: Mean residual covariance (McDonald & Mok, 1995). It is the mean difference between matrices of observed and reproduced item correlations.
- MADQ3: It is calculated by subtracting the model-predicted from the observed responses of the respondents and computing the average of the pairwise correlations of residuals (Yen, 1984).

Robitzsch, Kiefer, George, and Uenlue (2015) have proposed effect sizes for absolute model fit indices including MADcor, SRMSR and MADRESIDCOV (MADRCOV) which compare observed and predicted covariance (or correlations) of item pairs. The smaller an effect size, the better a model fits. The small effect sizes obtained in this study indicate a good model fit. It is important to note that there are not certain cutoffs to judge the absolute model fit indices yet.

We also compared the fit indices of 4 different models, including Generalized DINA (GDINA) (de la Torre, 2011), Deterministic Input, Noisy-And Gate Model (DINA) (Junker & Sijtsma, 2001), Deterministic Input, Noisy-Or Gate Model (DINO) (Templin & Henson, 2006), and the Additive Cognitive Diagnostic Model (ACDM) (de la Torre, 2011).

Among these models, DINO, and ACDM assume a compensatory relationship between the attributes while DINA is a non-compensatory model. GDINA, as a general model allows both compensatory and non-compensatory relationships between the attributes. The model fit indices are shown in Table 2.

Table 2*:* Relative Fit Indices

Model	Loglike	Deviance	AIC	BIC	AIC3	AICc	CAIC
GDINA	-60120.44	120240.9	120392.9	120940.9	120468.9	120394.1	121016.9
ACDM	-60355.92	120711.8	120843.8	121319.7	120909.8	120844.7	121385.7
DINO	-60660.77	121321.5	121433.5	121837.3	121489.5	121434.2	121893.3
DINA	-60678.07	121356.1	121468.1	121871.9	121524.1	121468.8	121927.9

Table 2 shows the information criteria AIC, BIC, AIC3, sample size adjusted AIC (AICc) and consistent AIC (CAIC) for the CDM models compared. The model with the smallest information criteria is the most preferable. According to the indices, the GDINA model fits the data best. The ACDM is the next best model. According to the underlying assumptions of the GDINA model which allows both compensatory and non-compensatory relationships between the attributes and also by taking the fact into account that ACDM is a compensatory model, we conclude that subskills underlying the reading comprehension ability interact in a compensatory manner and should be modeled with compensatory models such as ACDM. This finding is line whit the results of previous studies (Ravand & Robitzsch, 2015; Lee & Sawaki, 2009a; Li, Hunter, & Lei, 2015).

Model fit at item level

With general CDM models we can fit different models for each multi-attribute item within a test and select the best fitting model for each item (de la Torre & Lee, 2013, Ravand, 2016). Using the Wald test suggested by de la Torre and Lee (2013) we compared the fit of G-DINA, at item level, against that of other models. The results of the Wald test, using GDINA package (Ma, de la Tore, & Sorrel 2018), showed that for the 18 multi-attribute items LLM fits three items and RRUM fits two items and GDINA fits the rest of the items. However, the likelihood ratio test showed that the reduced model where items are allowed to pick their own model does not fit as good as the saturated model where GDINA is imposed on all the items, χ^2 (2)= 124.51, p<.001. Therefore, the GDINA is preferred for all the items. Note that a reduced CDM is suggested only for five items out of 18 (2 items have one attribute) and we save only two parameters by assuming the reduced CDMs for the five items.

G-DINA Parameters

Table 3 provides the model parameters which contain useful information about the probabilities of giving correct answers to each item based on the mastery of required attributes. The results for items 1 and 2 are illustrated. The second column of Table 3 represents the attributes required by each of the items. For instance, to answer the item 1 correctly, test-takers should master attributes 1 and 5, i.e., 'making inferences' and 'evaluating response options'. The third column shows the attribute mastery patterns. 1s indicate mastery of the required attribute and 0s indicate non-mastery of it. For example A10 represents the situation that the first required attribute is mastered while the second one is not. The fourth column displays the probability of success on each item for each attribute combination pattern.

As it is shown in Table 3, a candidate who has not mastered any of these attributes (pattern A00) has 2% chance of giving a correct answer to item 1. Generally, the pattern A00 indicates the probability of guessing an item right without mastering any of the required attributes.

Table 3: G-DINA Parameters

Item no.	Required Attributes	Mastery Pattern	Probability
1	A1-A5	A00	0.02
1	A1-A5	A10	0.22
1	A1-A5	A01	0.10
1	A1-A5	A11	0.61
2	A2-A5	A00	0.02
2	A2-A5	A10	0.40
2	A2-A5	A01	0.05
2	A2-A5	A11	0.44

Note: A1 to A5 are Making Inferences, Extracting Explicit Information, Identifying Word Meaning from the Context, Identify Pronominal References, and Evaluating Response Options, respectively.

For the same item, the second row of the table indicates the pattern that a candidate has mastered attribute 1 but has not mastered attribute 5. In such a case the probability of success on the item is: 2%+22%=24%. Considering the third row of the table, the chances of success for the candidates who have mastered attribute 5 but have not mastered the attribute 1 (indicated by pattern A01) is: 2%+10%= 12%. Taking these two probabilities into account, we conclude that attribute 1, 'making inferences', discriminates more between its masters and non-masters in comparison to attribute 5, i.e., 'evaluating response options'. In other words, mastery of attribute 1 increases chances of success on the item more than mastery of attribute 5. According to the fourth row of Table 3, mastery of both attributes increases the chances of giving a correct answer to the first item 61%. The masters of both attributes have 61%+2%=63% chances of success and on item 1 if they do not *slip*. By slipping we mean the probability of giving a wrong answer to the item in spite of having mastered all the required attributes.

Class Probabilities

Next latent class probabilities, i.e., the skill mastery patterns into which respondents are assigned were examined. The latent class probabilities and their frequencies are illustrated in Table 4. The first column of Table 4 shows the numbers allocated to each of the possible latent classes. The number of latent classes varies as a function of the number of required attributes. In the present study the test-takers are classified into 2^5=32 different latent classes. The attribute profiles related to each latent class are illustrated in the second column of Table 4. The probabilities of each latent class are in the third column of the table and its frequency is shown in the fourth column.

Table 4 indicates that the first latent class with the attribute profile of (00000) has the highest class probability which is about 57%. It means that approximately 57% of the candidates of the Iranian University Entrance Examination are classified in this latent class. As the fourth column shows it is expected that 5768 candidates belong to the first latent class, the members of which, have not mastered any of the attributes.

Table 4: Class Probabilities

Latent Class	Skill Pattern	Probabilities	Class Expected Frequency
1	00000	0.5768	5768
2	10000	0.0129	129
3	01000	0	0
…..	…….	…….	…….
30	10111	0	0
31	01111	0.0343	343
32	11111	0.1078	1078

The second latent class with the highest probability is the one with the attribute profile of (11111), latent class 32, with the probability of approximately 10%. The candidates who belong to this latent class are expected to have mastered all of the attributes. Therefore in the present study the attribute profiles of (00000) and (11111) have the highest probability in comparison to other thirty classes with an expected total frequency of 6746 candidates.

Class Probabilities for Respondents

We also calculated the probability for each respondent belonging to any of the latent classes. The results for three candidates with three different response patterns and total raw scores of 0, 10, and 5 are illustrated in Table 5.

Table 5: Class Probabilities for Respondents

Latent class	Response Pattern		
	00000000000000000000	01111101011100100000	00001000011100000010
Class1	0.98	0.00	0.00
Class2	0.00	0.00	0.03
Class3	0.00	0.00	0.00
Class4	0.00	0.00	0.07
Class5	0.00	0.00	0.00
Class6	0.00	0.00	0.10
Class7	0.00	0.00	0.00
Class8	0.00	0.01	0.28
Class9	0.00	0.00	0.00

 ISSN 2476-5880

Class10	0.00	0.00	0.00
Class11	0.00	0.00	0.00
Class12	0.00	0.00	0.00
Class13	0.00	0.00	0.00
Class14	0.00	0.00	0.11
Class15	0.00	0.00	0.00
Class16	0.00	0.01	0.18
Class17	0.00	0.00	0.00
Class18	0.00	0.00	0.00
Class19	0.00	0.00	0.00
Class20	0.00	0.00	0.00
Class21	0.00	0.00	0.00
Class22	0.00	0.00	0.00
Class23	0.00	0.00	0.00
Class24	0.00	0.00	0.00
Class25	0.00	0.00	0.00
Class26	0.00	0.00	0.00
Class27	0.00	0.00	0.00
Class28	0.00	0.00	0.00
Class29	0.00	0.00	0.00
Class30	0.00	0.00	0.00
Class31	0.00	0.00	0.00
Class32	0.00	0.98	0.23

Table 5 illustrates the probability for each respondent with his/ her response pattern belonging to a certain latent class. For instance, considering the candidate with a response pattern of (01111101011100100000), the probabilities for belonging to the latent classes 8, 16, and 32 are 0.01, 0.01, and 0.98 respectively. For such a candidate who has answered items 2,3,4,5,6,8,10,11, 12, and 15 correctly and could not answer the rest of the items successfully there is 98% chance of mastering all the attributes (belonging to latent class 32). In the same way the probability for such a candidate belonging to latent class 8 with the attribute profile of (11100) is only 0.01. To put it another way, there is only 1% chance that this candidate with such a response pattern has not mastered attributes four and five.

Attribute Mastery Probabilities

One of the most important issues in CDM is the attribute mastery probabilities. This helps us give feedbacks to the individual candidates on their strengths and weaknesses. The feedback is based on the probability that each respondent has mastered any of the attributes involved in answering the test items. A part of the results of analysis for this section is illustrated in Table 6.

The first column of Table 6 shows the ID number of each respondent. The second column represents the test-takers' response patterns and in the third column the attribute profiles of the respondents are shown. The next column shows the probability for each test-

 ISSN 2476-5880

taker belonging to the specified attribute profile assuming his/her response pattern. In columns 5 to 9, the probability that the respondent has mastered any of the attributes is given. For example, for respondent with the ID number 5504, there is 99% chance that s/he has mastered inference making, 93% chance that s/he has mastered extracting explicit information, 25%,14%, and 8% chance of mastering attributes 3, 4 and 5 respectively.

The results illustrated here also shows that two respondents with the same total scores essentially do not belong to the same latent classes and they might have different attribute mastery profiles. For instance, according to Table 6, both respondents 5504 and 9983 have answered 6 items successfully and, therefore they have the same total scores but they do not have the same mastery profiles, i.e., they have not mastered exactly the same reading comprehension subskills. This finding indicates the fact that leaners' strengths and weaknesses vary despite having the same total scores.

Table 6: Attribute Mastery Probabilities

Candidate No.	Pattern	attribute profile	Probability	Attribute1	Attribute2	Attribute3	Attribute4	Attribute5
1	0000000000000000000	00000	0.98	0.00	0.00	0.00	0.01	0.01
2140	1111011100 11111000 00	11111	0.98	0.99	0.99	0.99	0.99	0.98
5504	0000100100 01010001 10	11000	0.69	0.99	0.93	0.25	0.14	0.08
9983	0000001100 00101110 00	10110	0.65	0.97	0.33	0.96	0.98	0.21

The findings illustrated in this section about individual candidates are of great importance and have diagnostic value when we want to report fined-grained information about test-takers' reading comprehension ability. Based on the results provided here, each respondent can be informed of the problematic areas and subskills s/he needs to improve. Based on these results personal guidelines can be provided for each of the candidates when it comes to taking remedial actions as an attempt to enhance their own reading comprehension skill.

Attribute Difficulty

Attribute difficulty shows the percentage of examinees who have mastered each attributes (skill.prob1). We can say the more respondent who have mastered an attribute, the easier that attribute is. Therefore, based on the results illustrated in Table 7, the most difficult attributes for the candidates are 'making inferences' and 'evaluating response options'. Easier attributes are 'extracting explicit information', mastered by 26% of the candidates, followed by 'identifying word meaning from context'. The easiest attribute for the test-takers to master is 'identify pronominal references'.

Table 7: Attribute Difficulty

Attributes	skill.prob0	skill.prob1
Making inferences	0.76	0.24
Extracting explicit information	0.74	0.26
Identifying word meaning from context	0.70	0.30
Identify pronominal references	0.64	0.36
Evaluating response options	0.76	0.24

Note. skill.prob0 = probability of not mastering the attribute; skill.prob1 = probability of mastering the attribute.

Different levels of difficulty for reading comprehension attributes obtained from the model may be interpreted as evidence for existence of a hierarchy of difficulty for reading comprehension subskills. This is in line with other researchers findings (Grabe & Stoller, 2002; Lumley, 1993; Baghaei & Ravand, 2015; Ravand, 2016). Believing in such a hierarchy, teachers and learners should invest time and energy on each subskill in accordance to its level of difficulty.

Correlation between the Attributes

Table 8 depicts the corellations between the given attributes. The strongest corellation is between attributes 3 and 4, namely 'identifying word meaning from context' and 'identify pronominal references'. We conclude that the cognitive processes, which are tapped by these two attributes, are more or less simillar to each other. Perhaps these two attributes can be merged into one attribute. The weakest corellation is between attributes 1 and 5, 'making inferences' and 'evaluating response options'. We conclude that these two attributes have cognitively much less in common and they activate different undelying processes in respondents' minds.

Table 8: Attributes Correlations

Attribute	Att.1	Att.2	Att.3	Att.4	Att.5

Att.1	1	0.87	0.89	0.73	0.44
Att.2		1	0.82	0.82	0.85
Att.3			1	0.95	0.71
Att.4				1	0.93
Att.5					1

5. Discussion

Implementation of cognitive diagnostic modeling in educational measurement and cognitive psychology has enabled researchers to provide fine-grained information about test-takers performance on a variety of tests. The information derived from CDMs is valuable to the stakeholders including test-takers, educators, curriculum developers and syllabus designers who do not have access to this type of information through traditional scoring systems.

Retrofitting the CDMs to the existing non-diagnostic tests has become prevalent and has affected the score reporting systems. Employing CDMs, researchers can report the results, not only at the level of total row scores, which is useful for comparing examinees performance on a given test, but they can also provide stakeholders with skill mastery profiles which shed light on individual test-takers' strengths and weaknesses in a desired skill.

In this study we investigated the implementation of cognitive diagnostic modeling to provide fined-grained information about IUEE candidates' reading comprehension skill. The probability of giving correct answers to the reading comprehension items varies based on different attribute mastery profiles. The results showed that some attributes discriminates more between their masters and non-masters in comparison to other attributes. Being aware of this fact, the learners who try to get prepared for the IUEE may decide to focus more on the subskills whose mastery increase chances of success on the reading items more than other subskills.

We also observed how the GDINA model classified the respondents into different latent classes based on their attribute mastery profiles. Results showed that the mastery profiles of (00000) and (11111), which are called *flat profiles,* have the highest probability and the most frequency. This finding is in line with the results of other CDM studies (Lee & Sawaki, 2009b; Li, 2011; Ravand, Barati & Widhiarso, 2013; Ravand, 2016).

The latent class (00000) represents the mastery profiles of non-masters of all attributes and (11111) represents the profiles of masters of all 5 attributes. In the literature, the high frequency of the two flat profiles is considered as evidence for unidimensionality of the test. Test-takers who belong to one of the flat profiles, either have mastered or have not mastered the construt as a whole integrated skill. For instance, a test-taker who has mastered reading comprehension, has mastered all of the pre-assumed subskills which constitute the construct as a whole, and consequently the subskills cannot be separated from each other.

An important finding of the analysis, is the relatively high frequency of non-masters of all attributes. Findings indicate that most of the candidates have not mastered any of the attributes required for answering reading comprehension items. This finding is alarming as it indicates the poor performance of the educational system and its inability to improve the reading comprehension ability of the majority of the students.

We also checked the class probabilities for each respondent. rResults show that how individual respondents with specific response patterns are categorized in different latent classes. It is imporrtant to note that in traditional scoring systems, usually a single total score is reported which is used to compare test-takers' performances with each other. In such a system, we assume that two candidates with the same total score almost have the same knowledge of a desired construct. However, by implementation of CDMs, in the current study, we illustrated that two candidates with the same total score had different skill mastery profiles and they belonged to different latent classes which indicates that they do not have the same knowledge of the given construct (i.e. reading comprehension); thus traditional scoring systems do not provide a reliable criterion to precisely judge about test-takers knowledge.

To provide more fined-grained information about individual test-takers we investigated their attribute mastery probabilities. We provided feedback for individual respondents on the extent to which they have mastered each of the required attributes. The information here is particularly useful for autonomous learners who desire to enhance their own reading comprehension ability based on diagnostic feedbacks.

Findings also enabled us to draw a hierarchy of difficulty for reading comprehension attributes. This hierarchy of difficulty sets milestones for learners who ought to master the reading comprehension subskills step by step to improve their performance on the tests. The results of the present study showed that among the five attributes, 'making inferences' and 'evaluatiing response options' were the most difficult attribute for the candidates to master. The examinees need to accomplish a chain of complicated cognitive processes successfully when they want to make inferences. In other words, inference making involves higher level processing of the information in the text (Grabe, 2009; Harding, Alderson, & Brunfaut, 2015) and, therefore, as a complex attribute, it is difficult to master (Hammadou, 1991; Long, Seely, Oppy, & Golding, 1996; Hosenfield, 1977; Kim, 2014). In the same way, examinees need to undertake complicated cognitive processes when they need to use the attriute of 'evaluating response options' which makes it difficult for them to master this attribute as well.

By taking the valuable information obtained from CDMs into account, we can provide teachers and students with precise information about students' performance on a given test. Therefore, it is time to shift from traditional scoring system to the scores provided by cognitive diagnostic assessment so that the learners know about their status of a given construct in more detail. In a learner-centered teaching approach, having learners' mastery profiles the teacher can tailor the instruction for individual learners based on their strengths and weaknesses.

An interesting area for further research can be the effect of test-type (e.g. multiple choice items, gap filling, open ended questions) on the diagnosis power of CDMs. Researchers can also investigate how the examinees' characteristics (e.g. gender, age, level of language proficiency) affect the reliability of results obtained from CDMs. What is more, there is not any standard method of Q-matrix development yet. For future studies, the researchers should focus on enhancing methods of Q-matrix specification. A practical method can be developing tests with diagnostic purpose instead of retrofitting existing tests. In case of diagnostic tests, the item-attribute relations are determined in advance, therefore the Q-matrix specification is more precise.

Acknowledgements: Purya Baghaei was supported by the Alexander von Humboldt Stiftung via the Group Linkage Programme.

6. References

Alderson, J. (2000). *Assessing reading.* New York: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics,* 36, 236-260.

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York: Routledge.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: OUP.

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.

de La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: development and applications. *Journal of Educational Measurement, 45*, 343–362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130.

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

de la Torre, J., & Chiu, C.-Y. (2010, April). *A general empirical method of Q-matrix validation*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*, 355–373.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*, 227-249.

DiBello, L. V., & Stout, W. (2007). Guest editor's introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*, 285–291.

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics. *Psychometrics, 26,* 979-1030. Amsterdam, The Netherlands: Elsevier.

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: Erlbaum.

Doornik, J. A. (2007). *Object-oriented matrix programming using Ox* (6th Ed.). London, England: Timberlake Consultants Press.

Effatpanah, F., Baghaei, P., Boori, A. (2019). Diagnosing EFL learners' writing ability: A Diagnostic Classification Modeling Analysis. *Language Testing in Asia, 9*:12.

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*, 1-28.

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

García, P. E., Olea, J., & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema, 26*, 372-377.

Goh, C.C., & Aryadoust, V. (2014). Examining the notion of listening subskill divisibility and its implications for second language listening. *The International Journal of Listening, 28*, 1–25.

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.

Haberman. S., & von Davier, M. (2006). Some notes on models for cognitively based skills diagnosis. In C. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (pp. 1031–1038). Amsterdam: Elsevier.

Hammadou, J. (1991). Interrelationships among prior knowledge, inference, and language proficiency in foreign language reading. *The Modern Language Journal*, *75*, 27–38.

Harding, L., Alderson, C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: elaborating on diagnostic principles. *Language Testing*.

Hartz, S. M. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilites: Blending theory with practicality* (Doctoral dissertation). University of Illinois, Urbana-Champaign, IL.

Hartz, S., Roussos, L., & Stout, W. (2002). Skills diagnosis theory & practice (ETS document prepared by ETS External Diagnostic Research Team).

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*, 92-100.

Hosenfield, C. (1977). Learning about learning: Discovering our students' strategies. *Foreign Language Annals*, *9*, 117–129.

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

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, Urbana-Champaign, IL.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing, 26,* 31-73.

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.

Kim, A. Y. A. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*.

Kim, A. Y. A. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, 32, 227-258.

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. DOI: 10.1080/15434300903079562

Lee, Y. W., & Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*, 172–189.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge: Cambridge University Press.

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.

Li, H., Hunter, C. V., & Lei, P. W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*. 0265532215590848.

Long, D. L., Seely, M. R., Oppy, B. J., & Golding, J. M. (1996). The role of inferential processing in reading ability. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text*, 189-214. Mahwah, NJ: Lawrence Erlbaum.

 ISSN 2476-5880

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, *10*, 211-234.

Ma, W., de la Tore, J., & Sorrel, M. (2018). GDINA: The Generalized DINA Model Framework. R Package version 2.1.15. https://cran.r project.org/web/packages/GDINA/index.html.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). *Measurement: Interdisciplinary Research and Perspectives, 11*, 71-137.

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*, 23-40. doi: 10.1207/s15327906mbr32-001

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

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment Research and Evaluation*, 20, 1-12.

Ravand, H., Barati, H., & Widhiarso, W. (2013). Exploring diagnostic capacity of a high stakes reading comprehension test: A pedagogical demonstration. *Iranian Journal of Language Testing*, *3*, 11-37.

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent development, practical issues and prospects. *International Journal of Testing*, *20*, 24-56.

Ravand, H., Baghaei, P., & Doebler, P. (2020). Examining parameter invariance in a general diagnostic classification model. *Frontiers in Psychology*, *10*: 2930.

Ravand, H., and Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension. *Educational Psychology, 38*, 1255–1277.

Robitzsch, A., Kiefer, T., George, A. C. & Uenlue, A. (2015). CDM: Cognitive Diagnosis Modeling. R package version 4.5-0. http://CRAN.R-project.org/package=CDM

Rupp, A. A. & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78-96.

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

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

von Davier, M. (2005). mdltm [Computer software]. Princeton, NJ: Educational Testing Service.

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, 67, 49–71.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8.*