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

The Construction and Validation of a Q-matrix for Cognitive Diagnostic Analysis: The Case of the Reading Comprehension Section of the IAUEPT

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

Abstract
Cognitive diagnostic models (CDMs) have received sustained attention in educational settings because they can be used to operationalize formative assessment to provide diagnostic feedback and inform instruction. A large number of CDMs have been developed over the past few years. An important component of all CDMs is a Q-matrix that specifies a particular hypothesis about the relationship between each test item and its required attributes. The purpose of this study was to construct and validate a Q-matrix for the reading comprehension section of the Islamic Azad University English Proficiency Test (IAUEPT), as an advanced English placement test designed to measure the language ability of Ph.D. candidates who tend to pursue their studies in the IAU. To achieve this, using item responses of 1152 candidates to twenty items of the reading section of the test, an initial Q-matrix was constructed based on theories and models of second/foreign language (L2) reading comprehension, previous applications of CDMs on L2 reading comprehension, and brainstorming and consensus of five content experts. Then, the initial Q-matrix was empirically validated using the method proposed by de la Torre and Chiu (2016) and checking mesa plots, and heatmap plot. Five attributes were derived for the reading comprehension section: vocabulary, grammar, making an inference, understanding specific information, and identifying explicit information. Finally, the analysis of the Generalized Deterministic Inputs, Noisy "and" Gate (GDINA) regarding absolute fit at the item- and test-level as well as three residual-based statistics showed the accuracy of the Q-matrix and a perfect model-data fit.

Keywords: Cognitive Diagnostic Models (CDMs); GDINA; Islamic Azad University English Proficiency Test (IAUEPT); Q-matrix, Reading comprehension attributes

1. Introduction

A ubiquitous concept in educational testing, especially language testing and assessment, is standardized high-stakes testing, which has important consequences for test takers. Most high-stakes tests are largely developed based on the conventional educational psychometric frameworks such as item response theory (IRT), and unidimensional item response models are typically used to analyze test scores. Although scores obtained from this framework are used to scale and rank test takers along a single latent proficiency continuum, these assessments consist of limited fine-grained information for practical instructional settings to identify the strengths and weaknesses of test takers (de la Torre, 2009). As argued by Pellegrino et al.

[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

(1999), assessments should provide information that is "interpretative, diagnostic, highly informative, and potentially prescriptive" (p. 335).

Along the same lines, cognitive diagnostic models (CDMs), also called diagnostic classification models (DCMs; Rupp & Templin, 2008), have received sustained attention in educational contexts because they can be used to operationalize formative assessment to provide diagnostic feedback and inform instruction. According to Rupp and Templin (2008), CDMs are "probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modeling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables" (p. 226). CDMs decompose items and tasks into different knowledge, processes, and strategies examinees require to have mastered to be able to correctly answer a set of test items (Embretson, 1983). In CDMs, the diverse elements of a cognitive domain, which are categorical latent variables, are called attributes, also known as *skills, sub-skills, knowledge, abilities, processes,* and *strategies*. In this study, these terms are used interchangeably. As Birenbaum et al. (1993) defined, attributes are any "procedures, skills, or knowledge a student must possess in order to successfully complete the target task" (p. 443). In effect, attributes are domain-specific knowledge required to show mastery in a particular cognitive domain (Leighton & Gierl, 2007). For example, reading comprehension is a general cognitive domain which involves several attributes such as grammar, vocabulary, making an inference, identifying explicit information, and understanding specific information, etc. To comprehend a text and perform successfully on a set of test items, readers should have mastered these attributes. Decomposing items and tasks make CDMs produce diagnostic profiles based on the non-mastery/mastery of each required attribute (Kunina-Habenicht et al., 2009).

Over the past few decades, a large number of CDMs have been proposed such as the rule space (RSM; Tatsuoka, 1995), the Deterministic Inputs, Noisy And Gate (DINA; Junker & Sijtsma, 2001), the attribute hierarchy method (AHM; Leighton et al., 2004), the deterministic inputs, noisy "or" gate (DINO; Templin & Henson, 2006), the reduced reparameterized unified model (RRUM or fusion model; Hartz, 2002), the compensatory RUM (C-RUM; de la Torre, 2011), linear logistic model (LLM; Maris, 1999), the additive CDM (A-CDM; de la Torre, 2011), the generalized deterministic, inputs, noisy "and" gate (GDINA; de la Torre, 2011), the general diagnostic model (GDM; von Davier, 2008), and the log-linear cognitive diagnosis model (LCDM; Henson et al., 2008). A great deal of research has already been conducted to use CDMs on different L2 skills and components to design true diagnostic tests (Liu et al., 2013; Ranjbaran & Alavi, 2017; Shafipoor et al., 2021) or retrofit the existing non-diagnostic tests (Aryadoust, 2018; Buck & Tatsuoka, 1998; Buck et al., 1997; Dong et al., 2021; Effatpanah, 2019; Effatpanah et al., 2019; Jang, 2009; Javidanmehr & Anani Sarab, 2019; Kim, 2015; Li, 2011; Li & Suen, 2013; Mehrazmay et al., 2021; Ravand, 2016; Ravand et al., 2020; Sawaki et al., 2009; Yi, 2017a), and the results have shown the usefulness of CDMs in diagnosing examinees' ability in different cognitive domains and providing diagnostic feedback.

Table 1.
Sample Q-matrix

Items	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	0	0	0
2	1	1	0	0
3	0	0	1	1
…	…	…	…	…

An important component of all CDMs is an incidence matrix, known as Q-matrix (Tatsuoka, 1983), which is the first step in working with CDMs. The nature and number of attributes measured by a test should be determined (Li & Suen, 2013). Similar to factor analysis, a Q-matrix is the loading structure of CDMs in which a particular hypothesis, based on a substantive theory of a construct, about the relationship between each test item and its required attributes, is specified. In a Q-matrix, each row indicates an item, and each column represents an attribute. Let's consider a test that taps *J* attributes, each of the test item *I* needs several attributes to be correctly responded. Such item-attribute relationships are collected into a $J \times I$ matrix, $Q = \{q_{ji}\}$, where $i = 1, \ldots, I$ and $j = 1, \ldots J$. If a test item requires one or more attributes, 1s are used to show the requirement; otherwise, 0s are used to show the item does not require the attributes. As an illustration, Table 1 presents an example for an imaginary test which taps four attributes. As can be seen, Item 1 requires only Attribute 1; Item 2 requires Attributes 1 and 2; and Attributes 3 and 4 are required for Item 3. Ravand and Baghaei (2019, p. 15) maintained that some considerations should be considered in the Q-matrix construction process: (1) Correct specification of the Q-matrix: What attributes each item measures should be accurately specified, (2) Design of the Q-matrix: What is the configuration of the attributes in the Q-matrix, and (3) The grain size of the attributes: How finely the attributes should be specified (For more information, see Ravand & Baghaei, 2020 and Li & Suen, 2013).

Researchers have proposed a variety of approaches for identifying attributes involved in a test to construct a Q-matrix. The approaches are content analysis of the test items, existing test specifications, content domain theories, review of the relevant literature, a panel of experts, dimensionality analysis, eye-tracking research, and think-aloud protocols (Embretson, 1991; Gao & Rogers, 2007; Jang, 2009; Leighton et al.,, 2004; Sawaki et al., 2009). Research has shown that proper specifications of attributes that underlie a set of given test items or tasks and their theoretically-sound relationships with items maximize the quality of cognitive diagnostic assessment (CDA; Lee & Sawaki, 2009a).

While there are several strategies for determining attributes, the Q-matrix construction process is typically carried out subjectively. This subjectivity of Q-matrix development might result in the presence of misspecifications in the Q-matrix that affect the model parameters, the accuracy of classifications, and ultimately invalid inferences (Chiu, 2013; de la Torre & Chiu, 2016; Kunina-Habenicht et al., 2012; Madison & Bradshaw, 2015). Consequently, a large number of Q-matrix validation methods have been developed to detect and modify misspecifications in Q-matrices (e.g., Barnes, 2010; Cai et al., 2018; Chiu, 2013; de la Torre, 2008; de la Torre et al., 2022; de la Torre & Chiu, 2016; Kang et al., 2019; Li et al., 2021; Ma & de la Torre, 2020; Nájera et al., 2019; Nájera, et al., 2020; Wang et al., 2018; Yu & Cheng,

2020, to name a few). Among these methods, the procedure suggested by de la Torre and Chiu (2016) is the most commonly used method for Q-matrix validation. The method firstly requires fitting the GDINA as a general model to the data. This method "proposes a discrimination index that can be used with a wide class of CDM subsumed by the GDINA model to empirically validate the Q-matrix specifications by identifying and replacing misspecified entries in the Q-matrix" without changing correct entries (de la Torre & Chiu, 2016, p. 253). The rationale behind this method is that a correct *q*-vector will differentiate the different latent groups for that item in terms of the probability of success; however, more homogeneous probabilities of success across the specified latent groups will be the result of a misspecified *q*-vector.

2. The Generalized Deterministic Inputs, Noisy "and" Gate (GDINA)

The Generalized Deterministic Inputs, Noisy "and" Gate (GDINA; de la Torre, 2011) is a saturated and general model which is considered as the generalization of the DINA model. GDINA model allows the presence of different relationships between attributes involved in the test (e.g., non-compensatory and compensatory). Just like the Analysis of Variance (ANOVA), in its saturated form, the GDINA includes all possible main and interaction effects. The GDINA relaxes the conjunctive assumption of the DINA model which classifies examinees into two groups for each item; however, the GDINA classifies individuals into $2^{k_j^*}$ latent groups, where k_j^* is the number of required attributes for item *j*. In fact, each group has its own probability of success. The probability of getting an item right for an individual with a skill pattern $\alpha_{lj}^*$is a function of the main effects and all the possible interaction effects among the k_j^* required skills for item *j:*

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

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 $\alpha_{\boldsymbol{k}'}$, which shows the change in the probability of a correct response due to the mastery of both α_k and $\alpha_{\boldsymbol{k}'}$; $\delta_{j12\ldots k_j^*}$ is the highest-order interaction effect due to $\alpha_{1,\ \ldots,}\alpha_{k_j^*}$, which indicates 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). As argued by de la Torre (2011), when appropriate constraints are imposed on the parameterization of the general models including the GDINA, many constrained CDMs, such as the DINA, DINO, A-CDM, LLM, and RRUM, can be derived from general models.

3. The Present Study

Overall, the purpose of the present study was to construct and empirically validate a Q-matrix for the reading comprehension section of the Islamic Azad University English Proficiency Test (IAUEPT), as an advanced English placement test designed to measure language ability of Ph.D. candidates who tend to pursue their studies in the IAU. The following research questions were posed for the study:

1- What are the major underlying L2 reading processes or attributes involved in successfully completing the reading comprehension section of the IAUEPT?
2- Does the GDINA model fit the reading comprehension section of the IAUEPT based on the final Q-matrix empirically validated?

4. Data for the Study

Data for this study were obtained from the April 2019 administration of the IAUEPT to 1152 candidates. IAUEPT is an obligatory language proficiency test for all Ph.D. students in various fields of study who pursue their studies in the IAU as a requirement for graduation. It is designed, developed, and administered by the IAU Testing Centre. The test is administered almost every month in eleven cities in Iran. The test consists of 100 items in three sections: vocabulary, grammar, and reading comprehension. Test takers should answer all the questions within 140 minutes. The reading comprehension section, which is the focus of this study, consists of 35 multiple-choice items in two parts: the first part comprises two reading passages of different lengths with a total of 20 items, and the second part is a multiple-choice cloze test comprising a passage with 15 gaps (e.g., 15 multiple-choice items).

For this study, only 20 items of the reading comprehension section of the test were analyzed. The reading section comprises two reading passages, each including 10 items.
The first passage includes a 486-word text with a readability score of 84.8 on Flesch Reading Ease Score scale and 6.5 on Gunning Fox scale; it was assessed as easy to read. The second the passage consists of a 196-word text with a readability score of 51.8 on the Flesch Reading Ease Score scale and 12.9 on Gunning Fox scale; it was assessed as hard to read. The total score ranged from 0 to 19 with a mean of 7.21 and a standard deviation of 3.28. Unfortunately, demographic information of the candidates, such as gender, age, major, etc., is unavailable. Using Cronbach alpha (α), the reliability coefficients of the test were calculated, and the result showed a value of 0.63, suggesting a moderate internal consistency.

Also participating were five experienced EFL (English as Foreign Language) university instructors. They served as content experts to specify the possible attributes measured by the reading comprehension items of the test. All of the instructors were non-native English speakers, knowing English as a foreign language and Persian as their native language. They had Ph.D. degrees in English Language Teaching (ELT) with at least ten years of teaching and assessing reading comprehension. They were asked to participate in the Q-matrix (Tatsuoka, 1983) construction phase of the study to stipulate the relationship between each item and its required attributes. A 2-hour training session was held to train experts or coders to know how to specify item-attribute relationships.

5. Data Analysis

For the purpose of this study, several steps were taken to construct and empirically validate a Q-matrix for the reading comprehension section of the IAUEPT. In the first step, an initial Q-matrix was developed according to theories and models of L2 reading comprehension, previous applications of CDMs on L2 reading comprehension, and brainstorming and consensus of five content experts, who specified the relationships between each item and its requisite attributes. In the second step, the procedure proposed by de la Torre and Chiu (2016)

was used to empirically validate the initial Q-matrix using the GDINA package version 2.8.8 (Ma et al., 2022) in the R statistical software (R Core Team, 2013). The modifications suggested by the software were meticulously analyzed by the five content experts and the researchers. The mesa plots for each item were also checked. Modifications were only applied and kept in the Q-matrix if they were in line with the substantive theory of L2 reading comprehension. Otherwise, the suggestions were discarded. Item-level fit statistics and the heatmap plot, indicating the dependency of item pairs, were further checked to investigate whether the modified Q-matrix is supported by the data. The fit of the GDINA model for both the initial and final Q-matrices was inspected.

Using marginal maximum likelihood estimation with EM (Expectation-Maximization) algorithm, the GDINA package can provide a framework for a series of cognitively diagnostic analyses for polytomous and dichotomous responses, conduct various Q-matrix validation methods, and calibrate various models (Ma et al., 2022). Different absolute fit indices were used to check the quality of the Q-matrix and explore the fit of the GDINA model to the data. The following absolute fit indices were evaluated to check model-data fit:

1- M2 (Chen & Thissen, 1997) is the mean difference between the model-predicted and observed response frequencies. Large values show dependencies between items. A significant *p*-value indicates the violation of item independency, and the model does not fit the data (Hu et al., 2016);

2- RMSEA2 (the root mean square error of approximation fit index for M2) is considered a measure of the difference between the observed covariance matrix and model-predicted covariance matrix for each degree of freedom (Chen, 2007, p. 467). It can range from 0 to 1. As suggested by Maydeu-Olivares and Joe (2014), values lower than 0.05 indicate a good fit. Hooper et al. (2008) considered models with RMSEA2 values < 0.06 as an adequate fit;

3- The standardized root mean squared residual (SRMSR) is the square root of the difference between the model-expected correlations and observed correlations between all item pairs (Chen, 2007). According to Maydeu-Olivares (2013, p. 84), values below 0.05 show a substantively negligible amount of misfit. However, Hu and Bentler (1999) suggested that SRMSR is expected to be within the ideal range of 0 and 0.08.

In addition to the aforementioned fit indices, three residual-based statistics at the item-level were also examined (Chen et al. 2013, p. 126): (1) proportion correct (*p*) refers to the residual between the observed and predicted proportion correct of individuals' correct responses to a set of test items, (2) log-odds ratio (*l*) is the residual between the observed and predicted log-odds ratios of item pairs, and (3) transformed correlations (*r*) is the residual between the predicted and observed Fischer-transformed correlation of the item pairs. Lower values show a better model-data fit. As noted by Chen et al. (2013), if the model fits the data, the value of these residual-based statistics should be close to zero for all items. Values not significantly different from zero, as indicated by Bonferroni adjusted *p*-values > 0.05, show a well-fitting model.

6. Results

For this study, different methods were used to identify the attributes that the test takers should have possessed to be able to give correct answers to the reading comprehension test items. Li and Suen (2013) argued that the use of various sources of evidence for creating a Q-matrix can increase the reliability of the Q-matrix. L2 reading comprehension, as an interaction between the reader and the text, is a complicated cognitive process which calls upon decoding and lexico-grammatical knowledge (Grabe, 2009; Zhang, 2012). In this study, the researchers first consulted the literature on language ability and L2 reading comprehension models and taxonomies (Alderson, 2000; Alderson & Lukmani, 1989; Bachman, 1990; Bachman & Palmer, 1996; Birch, 2002; Cohen & Upton, 2006; Fletcher, 2006; Francis et al., 2006; Hughes, 2003; Jang, 2009; Lumley, 1993; Munby, 1978; Phakiti, 2007; Pressley & Afflerbach, 1995; Purpura, 2004; Urquhart & Weir, 1998). For example, in his reading comprehension model, Hughes (2003, p. 139) enumerates 20 reading attributes: (1) identify pronominal reference; (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 the 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 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*, *how*); and (20) make pragmatic inferences.

The second source for identifying reading attributes was the previous CDM studies on L2 reading comprehension (Buck et al., 1997; Chen & Chen, 2016; Gao, 2006; Jang, 2009; Javidanmehr & Anani Sarab, 2019; Lee & Sawaki, 2009b; Li, 2011; Li et al., 2015; Li & Suen, 2013; Mehrazmay et al., 2021; Mirzaei et al., 2020; Ranjbaran & Alavi, 2017; Ravand, 2016; Ravand et al., 2020; Ravand & Robitzsch, 2018; Sawaki et al., 2009; Yi, 2017b). For instance, Jang (2005) specified a set of nine attributes for the reading section of the TOEFL: context-dependent vocabulary, context-independent vocabulary, syntactic and semantic linking, negation, textually explicit information, summarizing, mapping contrasting ideas into a mental framework, inferencing, and textually implicit information. In another study, Li et al. (2015) identified four attributes for the reading section of the MELAB test including vocabulary, syntax, extracting explicit information, and understanding implicit information. Table 2 gives a list of reading attributes extracted from previous CDM studies. Although some of the attributes may not be applicable to the test used in this study, they help to select appropriate and relevant attributes.

Table 2.

Summary of Reading Attributes Identified in Previous CDM Studies on L2 Reading Comprehension

Studies	Extracted Attributes
Jang (2005)	- Context-dependent vocabulary - Context-independent vocabulary - Syntactic and semantic linking - Negation - Textually explicit information - Summarizing - Mapping contrasting ideas into a mental framework - Inferencing - Textually implicit information
Gao (2006)	- Recognize and determine the meaning of specific words or phrases - Understand sentence structure and sentence meaning using syntactic knowledge - Understand the relationship between sentences and the organization of the text-Speculate beyond the text - Analyze the function/purpose of communication using pragmatic knowledge - Identify the main idea, theme, or concept, and skim the text for gist - Locate the specific information requested in the question and scan the text for specific details - Draw inferences and conclusions based on information implicit in the text - Synthesize information presented in different sentences or parts of the text - Evaluate the alternative choices
Jang (2009)	- Deducing the meaning of a word or a phrase by searching and analyzing text and by using contextual clues appearing in the text. - Determine word meaning out of context with recourse to background knowledge - Comprehend relations between parts of text through lexical and grammatical cohesion devices within and across successive sentences without logical problems - Read expeditiously across sentences within a paragraph for literal meaning of portions of text. - Read selectively a paragraph or across paragraphs to recognize salient ideas paraphrased based on implicit information in text. - Skim through paragraphs and make propositional inferences about arguments or a writer's purpose with recourse to implicitly stated information or prior knowledge - Read carefully or expeditiously to locate relevant information in text and to determine which information is true or not true. - Analyze and evaluate relative importance of information in the text by distinguishing major ideas from supporting details. - Recognize major contrasts and arguments in the text whose rhetorical structure contains the relationships such as compare/contrast, cause/effect or alternative arguments and map them into mental framework

Sawaki et al. (2009), and Yi (2017b)	- Understanding word meaning - Understanding specific information - Connecting information - Synthesizing and organizing information
Li (2011)	- Vocabulary - Syntax - Extracting explicit information - Connecting and synthesizing - Making inferences
Li and Suen (2013)	- Vocabulary - Syntax - Extracting explicit information - Connecting and synthesizing - Making inferences
Li et al. (2015)	- Vocabulary - Syntax - Extracting explicit information - Understanding implicit information
Ravand (2016) **Ravand and Robitzsch (2018)**	- Reading for details - Reading for inference - Reading for main idea - Syntax - Vocabulary
Chen and Chen (2016)	- Identifying explicit information - Generalizing main ideas - Interpreting conceptual meanings - Making inferences - Evaluating and commenting - Understanding charts and graphs - Expressing in written forms
Ranjbaran and Alavi (2017)	- Determining word meaning from context - Determining word meaning out of context - Comprehending text-explicit info - Comprehending text-implicit info - Skimming - Summarizing - Inferencing - Applying background knowledge - Inferring major ideas or writers purpose
Javidanmehr and Anani Sarab (2019)	- Vocabulary knowledge - Local comprehension - Making inferences - Syntactic knowledge - Connecting and synthesizing
Mirzaei et al. (2020)	- Vocabulary knowledge - Grammatical knowledge - Making inferences - Extracting explicit information - Skimming for general information

	- Understanding specific information - Connecting information - Synthetizing information - Synthesizing
Mehrazmay et al. (2021)	- Vocabulary - Main idea - Details

After reviewing the relevant literature and previous studies on the application of CDMs on L2 reading comprehension, a list of reading attributes was provided and given to five university instructors as content experts to brainstorm and identify possible attributes measured by the test. As argued by Lee and Sawaki (2009a), when CDMs are retrofitted to existing non-diagnostic tests for which there is no detailed information on a cognitive model of task performance, "brainstorming about possible attributes that elaborate on an existing test specification might serve as a good point of departure" (p. 176). The five experts specified a set of five attributes: *vocabulary (VOC), grammar (GRM), making an inference (INF), understanding specific information (USI),* and *identifying explicit information (IEI)*. The experts were also trained for a session to learn how to code the identified attributes measured by each item. They read each item, discussed the relationship between each item and its required attribute, and specified the required attributes that underlie a given test item. Then, a common Q-matrix was constructed. Table 3 shows the initial Q-matrix.

Table 3.
Initial Q-matrix

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

As the Q-matrix construction was performed by the domain experts, some entries in the matrix might be misspecified due to subjectivity. Therefore, it is reasonable to use empirical Q-matrix validation methods to revise any possible misspecification in the matrix. The Q-matrix in the present study was empirically validated based on the procedure suggested by de la Torre and Chiu (2016) using the GDIINA package to detect and revise any specification errors. Before running the procedure, the quality of the initial Q-matrix was evaluated using a set of absolute fit indices (e.g., M2, RMSEA2, SRMSR) as well as three residual-based statistics, including proportion correct (*p*), log-odds ratio (*l*), and transformed correlation (*r*). As Table 4 demonstrates, the GDINA model produced unsatisfactory fit indices. The value of M2 is significant, indicating that the model does not fit the data.

Although the values of SRMSR and RMSEA2 are below 0.05, the upper bound confidence interval of RMSEA2 is beyond 0.05. Furthermore, Table 5 depicts the absolute item-level fit indices of the GDINA model. Using the Bonferroni correction, the significance level of a Z-score can be adjusted. For $\alpha = 0.01$, 4.17 is considered as the critical Z-score. According to Chen et al., (2013), a maximum Z-score greater than the critical Z-score indicates that the model does not fit the data. Table 5 shows that the GDINA model adequately fitted the data based on proportion correct values (e.g., Max $Z = 0.2169 < 4.17$; adjusted p-value > 0.05), whereas the adjusted *p*-values for transformed correlation and log odds ratio are significant. This could be due to the presence of some misspecifications in the Q-matrix (Chen et al., 2013; Sorrel et al., 2017). Therefore, all the above evidence suggested that the initial Q-matrix needed some revisions.

As the procedure proposed by de la Torre and Chiu (2016) identifies possible misspecifications and provides suggestions for revising the Q-matrix, a few modifications were suggested for the initial Q-matrix. Except from three cases for which the suggestion was to turn 1s into 0s (e.g., for items 5, 12, and 16), in other cases, the suggestion was the insertion of 1s into the Q-matrix. Because wholehearted acceptance of all the suggestions based on statistical analyses is very simplistic, the suggested elements for each item were carefully analyzed and examined by the experts, and only sensible suggestions were applied.

To better understand the suggestions, the mesa plot for each item was checked. The mesa plot (de la Torre & Ma, 2016) is a line chart, in which the *x*-axis is the *q*-vectors for different numbers of *K* attributes, and the *y*-axis is the corresponding proportion of variance accounted for (PVAF) associated with those *q*-vectors. The mesa plot is similar to the scree plot in factor analysis (Ma, 2019). The *q*-vectors are ordered from the lowest to the highest PVAF so that the *q*-vector with all the possible attributes specified will always show the highest discrimination index because "the specification of additional attributes leads to the differentiation among more latent groups, and so to a higher variability in the probabilities of success" (Nájera et al., 2019, p. 7). Therefore, the correct *q*-vector will be the simplest one with the least number of required attributes which can explain the large proportion of variance. In the mesa plot, the red dots are the original *q*-vectors. The *q*-vector on the edge of the mesa is likely to represent the best *q*-vector for the item (de la Torre & Ma, 2016). The cutoff value for PVAF is set at $\varepsilon(EPS) = 0.95$. Figure 1 illustrates the mesa plot for item 9 in which the original *q*-vector, indicated with a red dot, is below the value of 0.95. The *q*-vector [11001] indicates that the presence of three attributes, e.g., VOC, GRM, and IEI, is needed to account for 93% of the variance of the

probabilities of success. Although this value is still under the cutoff value for PVAF, it is on the edge of the mesa plot, so it can be the best *q*-vector for the item. For item 13 (Figure 1), the requirement of VOC and IEI was specified. The mesa plot shows that the addition of the GRM to the Q-matrix increases the value of PVAF, so the *q*-vector [11001] is appropriate for the item.

Furthermore, the heatmap plot was checked to inspect dependencies between item pairs as indicated through transformed correlations. Figure 2 presents the heatmap plots in which *x*- and *y*-axes are items, and the first and the last items are dropped on both axes (Ma, 2019). The shading area represents the Bonferroni adjusted *p*-values for all item pairs. Red squares show *p*-values lower than 0.05, indicating insufficient fit, and grey squares indicate *p*-values larger than 0.05, showing sufficient fit. As can be seen in Figure 2a, there are significant dependencies between some item pairs in the initial Q-matrix. However, applying the reasonably suggested modifications removed the dependency for the final Q-matrix (Figure 2b).

Figure 1.
Mesaplots for Two Items

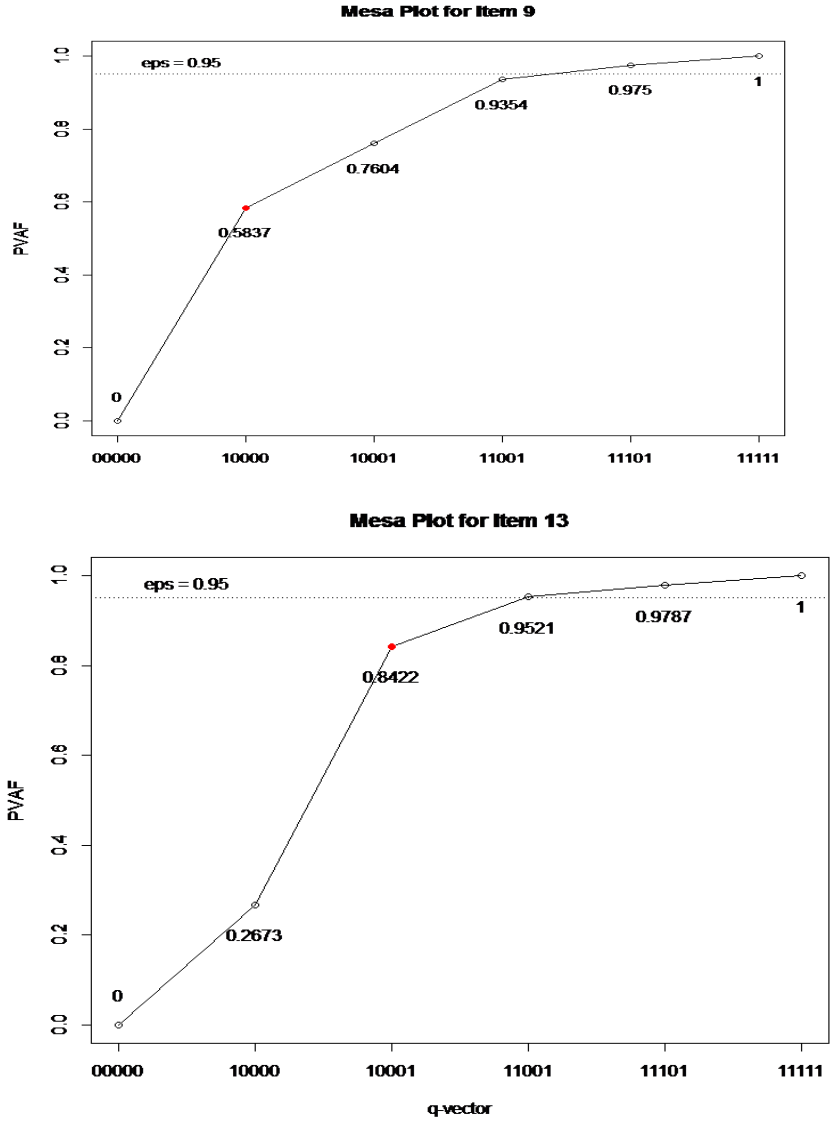
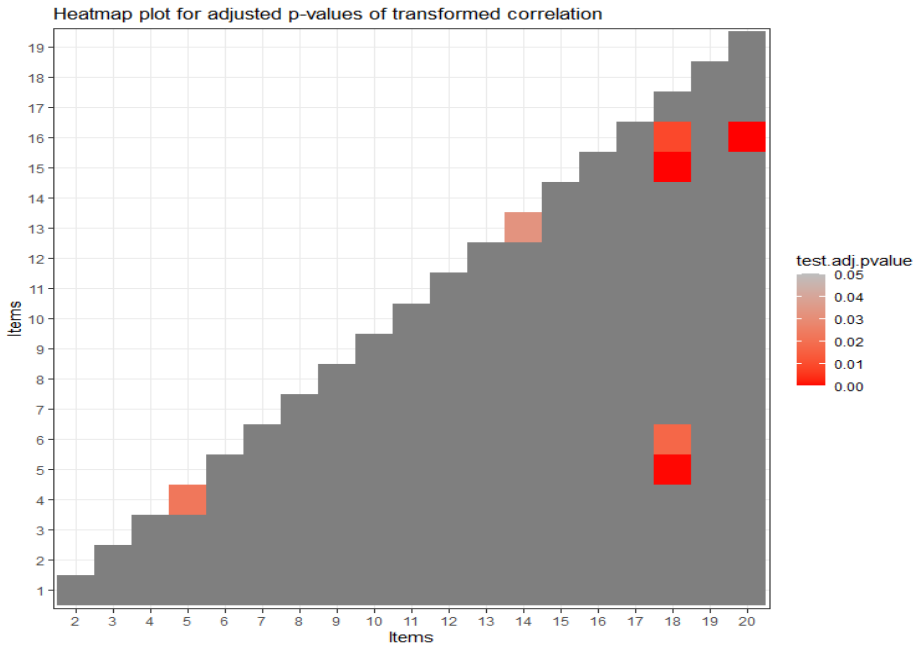
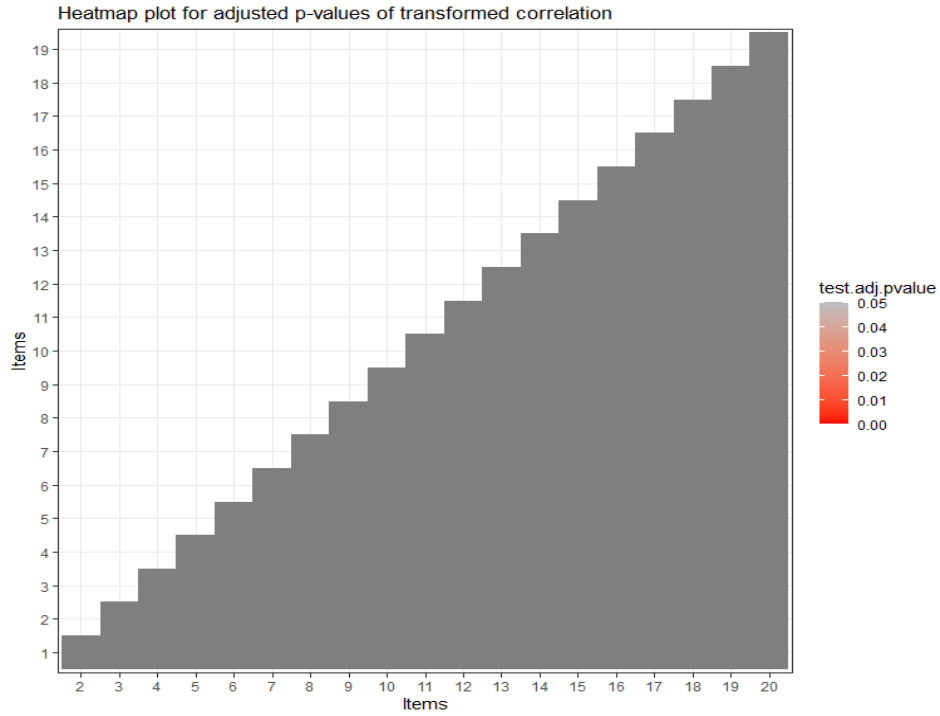


Figure 2.

Heatmap Plot of Adjusted p-values for the Initial and Final Q-matrices



(a)



(b)

As Tables 4 and 5 illustrate, by applying the sensible suggestions, the fit of the GDINA model improved. The values of relative fit statistics (e.g., -2log likelihood, AIC, and BIC) showed that the final Q-matrix has a better fit compared to the initial Q-matrix. The value of M_2 was 32 with a non-significant *p*-value (e.g., 0.25), suggesting a good fit of the model to the data. With regard to $RMSEA_2$, the value of the GDINA model (e.g., 0.0122) and its upper and lower bound were lower than 0.05. Also, in relation to SRMSR, the value is below the 0.05. Overall, the results of absolute fit statistics at both test and item-level revealed that the GDINA model has a perfect fit to the data. Table 6 shows the final Q-matrix.

Table 4.
GDINA Fit Indices for Initial and Final Q-matrices

Models	**M_2 (*p*-value)**	**Npar**	**RMSEA2**	**RMSEA2. CI1**	**RMSEA2. CI2**	**SRMSR**	**-2log likelihood**	**AIC**	**BIC**
Initial Q-matrix	264.2537 (0)	137	0.0477	0.0416	0.0539	0.0465	27033.29	27307.29	27999.04
Final Q-matrix	32 (0.25)	183	0.0122	0	0.0271	0.031	26694.81	27060.81	27984.83

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

Table 5.
GDINA Item Fit indices for Initial and Final Q-matrices

		mean[stats]	max[stats]	max[z.stats]	*p*-value	adj. *p*-value
Initial Q-matrix	**Proportion correct**	0.0011	0.0030	0.2169	0.8283	1
	Transformed correlation	0.0349	0.1858	6.2982	0.0000	0
	Log odds ratio	0.1606	0.7807	6.1571	0.0000	0
Final Q-matrix	**Proportion correct**	0.0012	0.0032	0.2204	0.8255	1
	Transformed correlation	0.0248	0.0994	3.3704	0.0008	0.14
	Log odds ratio	0.1135	0.4382	3.2206	0.0013	0.24

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

Table 6.
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

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

7. Discussion and Conclusion

An important first step in CDMs is the construction of a Q-matrix (Tatsuoka, 1983), which indicates the required attributes to correctly answer each item of a given test, that is, it specifies the particular theoretically based item-attribute relationships. The robustness of the Q-matrix has a great impact on cognitive diagnostic modeling and inferences drawn from the analysis of CDMs (Jang, 2009). There are various strategies for identifying attributes involved in successfully completing items of a test and developing a Q-matrix. The strategies are content analysis of test items, test specifications, content domain theories, literature review, a panel of experts, dimensionality analysis, eye-tracking research, and think-aloud protocols. However, these strategies are mostly subjective and "confirmatory in the sense that they assume that the proposed Q-matrix is known" (Nájera et al., 2019, p. 4). Studies have shown that specification errors in a Q-matrix can dramatically affect the estimation of model parameters and the precision of classifications (de la Torre & Chiu, 2016; Kunina-Habenicht et al., 2012; Madison & Bradshaw, 2015).

This study set out to construct and empirically validate a Q-matrix for the reading comprehension section of the IAUEPT. As argued by Li and Suen (2013), a sensible Q-matrix depends on evidence from multiple sources. To explore attributes measured by the IAUEPT, the researchers adopted a sequential combination of results from three sources. First, the relevant literature on L2 reading comprehension including reading models and taxonomies was reviewed to identify the exact nature of reading comprehension. Second, previous applications of CDMs on L2 reading comprehension were examined to know what attributes have already been used to construct CDMs and conduct CDM analysis. Third, five context experts were invited to first brainstorm the essential attributes required to give correct answers to reading items of the IAUEPT and then code the relationships between each item and its required attributes. An initial Q-matrix was constructed by integrating evidence from these sources. Five attributes were derived: vocabulary, grammar, making an inference, understanding specific information, and identifying explicit information.

A potential problem in building a Q-matrix is the subjective nature of the Q-matrix construction process. For this reason, many Q-matrix validation methods can be used to detect and modify Q-matrix misspecifications. The initial Q-matrix was empirically validated using the GDINA as a general model. The procedure suggested by de la Torre and Chiu (2016) was used to detect and revise misspecified entries in the Q-matrix. As wholeheartedly accepting statistical suggestions may jeopardize the soundness of the Q-matrix (Li & Suen, 2013), the suggested revisions were analyzed by the content experts and researchers to agree with both substantive theory and statistical modeling (Jang, 2009). The mesa plots for each item and the heatmap plot for identifying item pairs dependencies were also checked. After several rounds of refinements, a final Q-matrix was developed. Finally, the fit of the GDINA model at both test- and item-level as well as some residual-based statistics was examined, and the results indicated the perfect fit of the model to the data and robustness of the Q-matrix.

This study includes some limitations. First, although the Q-matrix developed in this study went through a prudent empirical Q-matrix validation procedure, it is not the only possible Q-matrix for the reading section of the IAUEPT. Future studies can take another approach for the appropriate grain size of attributes and use various strategies to construct a Q-matrix. However, Li and Suen (2013) highlighted that

> The more skills identified, the richer the diagnostic information that can be provided. However, including a large number of skills places a stress on the capacity of statistical modeling given the fixed length of the test. One important implication for test developers is, therefore, to keep a balance between the number of skills being measured and the number of items in the test; that is, more items should be included if more fine-grained diagnostic information is of interest. As suggested by Jang (2009), decisions about the grain size of the skills should be made by considering theoretical (construct representativeness), technical (the availability of test items), and practical (the purposes and context of using diagnostic feedback) factors. (p. 21)

Second, a non-diagnostic test was utilized in the present study. According to Gierl and Cui (2008), "a cognitive model would be developed first to specify the knowledge and skills evaluated on the test, and then items would be created to measure these specific cognitive skills" (p. 265). However, there are very few tests constructed based on a cognitive diagnostic purpose. The Q-matrices have thus been developed for existing non-diagnostic tests in a retrospective manner (Li & Suen, 2013). In other words, unlike a true CDM in which essential attributes are required a priori, retrofitting existing non-diagnostic tests requires careful test and attribute specifications. This retrofitting of existing tests may decrease the quality of CDA in providing fine-grained diagnostic information (Jang, 2009). Though, Lee and Sawaki (2009b) contend that retrofitting efforts can be considered a critical step in advancing cognitive diagnostic assessment of reading and are worthwhile for "examining the extent to which useful cognitive diagnostic information could be extracted from existing assessments before delving into an expensive, time-consuming process of designing a new cognitive diagnostic test" (p. 174).

References

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

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, *5*(2), 253-270.

Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore–Cambridge general certificate of education O-level: Application of DINA, DINO, GDINA, HO-DINA, and RRUM. *International Journal of Listening*, *35*(10), 1-24. https://doi.org/10.1080/10904018.2018.1500915

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

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

Barnes, T. (2010). Novel derivation and application of skill matrices: The Q-matrix method. In C. Ramero, S. Vemtora, M. Pechemizkiy, & R. S. J. de Baker (Eds.), *Handbook of educational data mining* (pp. 159-172). Boca Raton, FL: Chapman & Hall.

Birch, B. M. (2002). *English L2 reading: Getting to the bottom*. Routledge.

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

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119-157. https://doi.org/10.1191/026553298667688289

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

Cai, Y., Tu, D., & Ding, S. (2018). Theorems and methods of a complete Q-matrix with attribute hierarchies under restricted Q-matrix design. *Frontiers in Psychology*, *9*, 1413.

https://doi.org/10.3389/fpsyg.2018.01413

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

Chen J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement, 41*(4), 277-293. https://doi.org/10.1177/0146621616686021

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

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

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598-618. https://doi.org/10.1177/01466216134884

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks*. Princeton, NJ: ETS [TOEFL Monograph No. MS-33].

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*(4), 343-362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163-183. https://doi.org/10.1177/0146621608320523

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

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

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

de la Torre, J., Qiu, X. L., & Santos, K. C. (2022). An empirical Q-matrix validation method for the polytomous GDINA model. *Psychometrika, 87*(2), 693-724. https://doi.org/10.1007/s11336-021-09821-x

Dong, Y., Ma, X., Wang, C., & Gao, X. (2021). An optimal choice of cognitive diagnostic m for second language listening comprehension test. *Frontiers in Psychology, 12*, 1-12. https://doi.org/10.3389/fpsyg.2021.608320

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

Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability:

A diagnostic classification modeling analysis. *Language Testing in Asia, 9*(12), 1-23. https://doi.org/10.1186/s40468-019-0090-y

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197. https://doi.org/10.1037/0033-2909.93.1.179

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

Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading, 10*(3), 323-330. https://doi.org/10.1207/s1532799xssr1003_7

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading, 10*(3), 301-322. https://doi.org/10.1207/s1532799xssr1003_6

Gao, L. (2006). Toward a cognitive processing model of MELAB reading test item performance. *Spann Fellow Working Papers in Second or Foreign Language Assessment, 4*, 1-39. English Language Institute, University of Michigan, MI. Retrieved from https://michiganassessment.org/wpcontent/uploads/2020/02/20.02.pdf.Res_. TowardaCognitiveProcessingModelofMELABReadingTestItemPerformance.pdf

Gao, L., & Rodgers, T. (2007, April). *Cognitive-psychometric modeling of the MELAB reading items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research & Perspective, 6(4),* 263-268. https://doi.org/10.1080/15366360802497762

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

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.

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).

Hu, L., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

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

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

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

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing*, *26*(1), 31-73. https://doi.org/10.1177/0265532208097336

Javidanmehr, Z., & Anani Sarab, M. R. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the GDINA 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

Kang, C., Yang, Y., & Zeng, P. (2019). Q-matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement*, *43*(7), 527-542. https://doi.org/10.1177/0146621618813104

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

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*(2-3), 64-70. https://doi.org/10.1016/j.stueduc.2009.10.003

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

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

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

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

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

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

Li, H., & 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

Li, J., Mao, X, & Zhang, X. (2021). Q-matrix estimation (validation) methods for cognitive diagnosis. *Advances in Psychological Science, 29*(12), 2272-2280. URL: https://journal.psych.ac.cn/xlkxjz/EN/10.3724/SP.J.1042.2021.02272

Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., & Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30*(2), 152-172. https://doi.org/10.1007/s00357-013-9128-5

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

Ma, W. (2019). Cognitive diagnosis modeling using the GDINA R package. In M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 593-601). Switzerland: Springer Nature.

Ma, W., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 142-163. https://doi.org/10.1111/bmsp.12156

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

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75*(3), 491-511. https://doi.org/10.1177/0013164414539162

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

Mehrazmay, R., Ghonsooly, B., & de la Torre, J. (2021) Detecting Differential Item Functioning 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

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

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

Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical Q-matrix validation. *Educational and psychological measurement, 79*(4), 727-753. https://doi.org/10.1177/0013164418822700

Nájera, P., Sorrel, M. A., de la Torre, J., & Abad, F. J. (2020). Improving robustness in Q-matrix validation using an iterative and dynamic procedure. *Applied Psychological Measurement, 44*(6), 431-446. https://doi.org/10.1177/0146621620909904

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 307-353). Washington, DC: American Educational Research Association.

Phakiti, A. (2007). *Strategic competence and EFL reading test performance: A structural equation modeling approach*. Peter Lang, Frankfurt am Main.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge: 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. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782-799. https://doi.org/10.1177/0734282915623053

Ravand, H., & Baghaei, P. (2019). Diagnostic Classification Models: Recent Developments, Practical Issues, and Prospects. *International Journal of Testing*, 1-33. https://doi.org/10.1080/15305058.2019.1588278

Ravand, H., Baghaei, P., & Doebler, P. (2020). Examining Parameter Invariance in a General Diagnostic Classification Model. *Frontiersin Psychology*, 10(2930). https://doi.org/10.3389/fpsyg.2019.02930

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

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

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

Shafipoor, M., Ravand, H., & Maftoon, P. (2021). Test-level and item-level model fit comparison of General vs. specific diagnostic classification models: A case of True

DCM. *Language Testing in Asia, 11*(33), 1-20. https://doi.org/10.1186/s40468-021-00148-z

Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 13*(Suppl 1), 39-47. https://doi.org/10.1027/1614-2241/a000131

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

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

Templin, J. 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

Urquhart, S., & Weir, C. (1998). *Reading in a second language: Process, product, and practice*. Longman: Addison Wesley/Longman.

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

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement*, *42*(6), 446-459. https://doi.org/10.1177/0146621617752991

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

Yi, Y. (2017b). 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

Yu, X., & Cheng, Y. (2020). Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnostic assessment. *The British Journal of Mathematical and Statistical Psychology*, *73 Suppl 1*, 145-179. https://doi.org/10.1111/bmsp.12191

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