Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 11, No. 1, March 2021
ISSN 2476-5880

The Construction and Validation of a Q-matrix for a High-stakes Reading Comprehension Test: A G-DINA Study

Fateme Roohani Tonekaboni[1]*, Hamdollah Ravand [2], Reza Rezvani [3]

Received: 14 December 2020 Accepted: 12 February 2021

Abstract

Investigating the processes underlying test performance is a major source of data supporting the explanation inference in the validity argument (Chappelle, 2021). One way of modeling the cognitive processes underlying test performance is by constructing a Q-matrix, which is essentially about summarizing the attributes explaining test-takers' response behavior. The current study documents the construction and validation of a Q-matrix for a high stakes test of reading within a generalized-deterministic inputs, noisy "and" gate (G-DINA) model framework. To this end, the attributes underlying the 20 items of the reading comprehension test were specified through retrospective verbal reports and domain experts' judgments. In the ensuing stage, the Q-matrix thus developed along with item response data of 2625 test-takers were subjected to empirical analysis using the procedure suggested by de la Torre and Chiu (2016). Item-level results showed that, except for one item, the processes underlying the other items were captured by compensatory and additive models. This finding has significant implications for model selection for DCM practitioners.

Keywords: *cognitive diagnostic assessment; Q-matrix construction; Q-matrix validation; test reading comprehension*

1. Introduction

High stakes language tests often fail to provide test takers with diagnostic information that can be used to support learning (Afflerbach, 2004, 2016; International Literacy Association, 2017; Rajagopalan & Gordon, 2016). One approach to compensate for this pitfall is to complement such tests with more learning-friendly assessment approaches. Cognitive diagnostic assessment (CDA) is one such approach where instead of merely telling the test takers which items they have got wrong or right, the underlying test response processes are used to give diagnostic feedback that can inform and support further learning. More specifically, CDA, the offspring of the fields of education and cognitive psychology, is the process of obtaining a skill-based classification of an individuals' current latent knowledge status in a specific domain based on their observed responses to make finer-grained inferences and decisions for providing timely follow-up and support (Rupp & Templin, 2008;

[1]*Shahid Chamran University of Ahvaz, Iran, f_roohani_t@yahoo.com (corresponding author)*
[2]*Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran, ravand@vru.ac.ir*
[3]*Yasouj University, Yasouj, Iran, rrezvani@yu.ac.ir*

Rupp, Templin & Henson, 2010; de la Torre & Minchen, 2014). Although the ideal of CDA is realized when tests are founded upon a sound cognitive model at the design stage (Leighton & Gierl, 2007; Rupp & Templin, 2008; Rupp et al., 2010), CDA does have the potential to inform the cognitive potential of existing tests (Jang, 2009; Javidanmehr & Anani Sarab, 2017; Hemati & Baghaei Moghadam, 2020; Hemati, Baghaei, & Bemani, 2016; Kim, 2015; Lee & Sawaki, 2009a; Li & Suen, 2013; Li, Hunter & Lei, 2016; Liu, Huggings-Manley & Bulut, 2018; Ravand & Baghaei, 2020; Ravand, Barati, & Widhiarso, 2013; Rupp et al., 2010)that have not been designed based on an explicit cognitive theory.

Except for a few studies (e.g., Henson & Douglas, 2005) that sought to develop a test within a CDA framework at the design stage, namely true DCM studies, most high-stakes tests are not designed based on a clearly articulated cognitive model (DiBello, Roussos, & Stout, 2007; Sessoms & Henson, 2018; Ravand & Baghaei, 2020; Rupp & Templin, 2008). As a result, most CDA studies are either methodological for model development and refinement or retrofitting to existing non-diagnostic tests. Retrofitting CDA studies are basically for model demonstration or construct identification (Ravand & Baghaei, 2020). Despite efforts made to address the application of various DCMs (e.g., Jang, 2009; Kim, 2015; Lee & Sawaki, 2009b; Li et al., 2016; Ravand & Robitzsch, 2018; Yi, 2012), the construction and validation of Q-matrices in these studies have been treated subsidiary to the application of the models. The validity of a DCM study rests heavily on the quality of the Q-matrix that is input into the DCM analysis. Except for the trailblazing study by Lee and Sawaki (2009), few attempts have been made to delineate the specifics of Q-matrix construction and validation. Back in 2009, none of the software programs could deal with empirical validation of Q-matrices; therefore, Lee and Sawaki had to make do with qualitative analysis of a select group of the test taker responses to come up with the Q-matrix. The present study attempts to walk the readers through the specifics of qualitative and quantitative procedures taken to construct and validate the Q-matrix.

Hence, this study aims to identify the underlying attributes of the reading comprehension section of the high-stakes university entrance examination (UEE) Master of Arts (M.A.) exam, which serves a gate-keeping function to graduate English language programs in Iran. It is a standardized multiple-choice, speed test comprising general proficiency and subject matter knowledge. In this paper, besides substantive and empirical validation of the Q-matrix under a general DCM framework, the possibility of replacing G-DINA with simpler models is also investigated. Before discussing how this study was carried out, a brief review of literature pertaining to reading assessment is in order.

2. Review of Literature

2.1. The Construct of Reading

In both psychological and educational assessment, a crucial first step is to define the target construct that is to be measured. It follows that any assessment of L2 reading requires that reading be defined. Yet, although reading has a rather older tradition, research into the construct of reading is rather recent (Grabe & Jiang, 2013). That said, recent studies about human cognition have furthered our understanding of the constitutive components and

structure of the reading process. According to National Assessment Governing Board (NAEP; 2015), reading is defined as a dynamic cognitive process that involves "understanding written text, developing and interpreting meaning, and using meaning as appropriate to the type of text, purpose, and situation" (p. 2).

The NAEP's portrayal of reading as a dynamic, strategic, and goal-oriented process involving strategies, skills, prior knowledge, and the reader's purpose together with the anticipation of the types of reading assessment necessary to gauge student growth in reading across the school also shaped the nature of the construct in the Programme for International Student Assessment's (PISA) framework. Replacing the construct of reading with "reading literacy," PISA defines "reading literacy" as "understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2014, p.61), which is influenced by the reader, text, and task factors.

Despite the great influence of L1 reading models (i.e., bottom-up, top-down, and interactive models; Barnett,1989) on our understanding of ESL/EFL reading process, the peculiarities of the needs of ESL/EFL readers with varied linguistic and cultural knowledge of the English Language cannot be readily addressed within these models. It should also be noted that the borders between L1 and L2 reading are not necessarily clearly defined, as one of the great debates surrounding the construct of L2 reading is about whether it is an L1 ability that is simply carried over to the L2 or if it is part of the broad construct of L2 proficiency (Grabe & Stoller, 2013).

Therefore, componential models of the reading process (e.g., Bernhardt, 1986, 2010; Coady, 1979) have been developed with their specific focus "on the different types of components involved in reading such as conceptual abilities, process strategies, and background knowledge, rather than the process of reading" (Ghaith, 2018, p. 3). The main issues in applying these models to L2 reading have to do "with whether L2 reading is a developmental process and whether knowledge of different areas of reading comprehension can compensate for each other" (Ghaith, 2018, p. 3). Moreover, modified interactive models (Ghaith, 2018, p. 3) have been suggested to explain the L2 reading process (e.g., Dana & Hedgcock, 2009). These models see the ordinary interactive models self-contradictory "since the essential components of bottom-up processing (i.e., efficient automatic processing in working memory) are incompatible with the strong top-down controls because these controls are not automatic" (Ghaith, 2018, p. 3). Therefore, the role of bottom-up and top-down processes are respectively emphasized and minimized in the modified interactive models "on the assumption that activating prior knowledge or schematic resources may be time-consuming. As such, a reader may recognize words by perceiving information from graphemes, phoneme-grapheme correspondences, and spelling without employing schematic knowledge" (Ghaith, 2018, p. 3).

One perennial issue in assessment is that performance under testing conditions does not simulate real-world tasks. This issue of authenticity versus artificiality directly affects the nature of the L2 reading construct. Inspired by Afflerbach (2017), researchers in this study define test reading comprehension as the act of constructing meaning from text using required sub-skills/ attributes, strategies, and prior knowledge to answer high stakes test

questions. Magliano, Millis, Ozuru, and McNamara (2007) categorized reading comprehension assessment in the context of strategy interventions into two categories based on the goals such assessment is designed to achieve: general classification of readers and diagnosing readers' specific weakness or problemse Both of these categories can be touched on within the DCMs framework. A brief description of these models is brought in the following section.

2.2. Diagnostic Classification Models

DCMs as a family of latent class models (Wang, Shu, Shang, & Xu, 2015) provide a novel approach to analyzing test scores and conducting diagnosis through theoretical modeling and statistically examining test-takers' cognitive processes. These models seek to classify examinees as masters or non-masters on a set of test sub-skills/ attributes and provide more fine-grained diagnostic information about the quality of items and attributes measured by the items (DiBello et al., 2007; Lee & Sawaki, 2009a; Rupp et al., 2010). DCM application rests on a 2-way item by attribute Q-matrix with 1s and 0s indicating examinees' mastery or non-mastery of a certain attribute, respectively. Several DCMs have been developed and proposed in the literature whose selection hinges on a number of criteria such as identifiability of the model, interpretability of model parameters, interaction among attributes/sub-skills, i.e., assuming non-compensatory, compensatory, additive relations or structures among attributes, measurement scales of items and sub-skills, and the availability of the software program (Lee & Sawaki, 2009b, p.181). Some of these diagnostic classification models, which can be encompassed in general/ saturated models (e.g.,, G-DINA) are highly constrained like the deterministic inputs, noisy "and" gate (DINA; Junker& Sijitsma, 2001) and the deterministic inputs, noisy "or" gate (DINO; Templin & Henson, 2006) models; some enjoy the additive nature like the additive CDM (A-CDM; de la Torre, 2011), the linear logistic model (LLM; Maris, 1999), and the reduced reparameterized unified model (R-RUM; DiBello et al., 2007; Hartz, 2002). These models can be developed from G-DINA by imposing some constraints on the parameterization of general models and changing the link function. Moreover, changing the link function is shown (de la Torre, 2011) to result in other general models such as the log-linear cognitive diagnostic model (LCDM; Henson, Templin, & Willse, 2009) and general diagnostic model (GDM; von Davier, 2005).

G-DINA model

De la Torre (2011) proposed a general DCM, called the generalized deterministic inputs "and" gate (G-DINA) model with the identity link enjoying all possible item effects, for instance, the intercept or guessing, main effects, and interaction effects between all possible combinations of attributes. The probability of correctly answering an item requiring two attributes α_1 and $_2$of for G-DINA in its saturated form can be written as follows:

$$P\ (X_j = 1|\ \alpha_{1,}\ \alpha_2) = \delta j_0 + \ _{j1}\delta\alpha_1 + \delta_{j2}\alpha_2 + \delta_{j12}\alpha_1\alpha_2$$

In the equation, δj_0 as the intercept for item j shows the baseline probability, which is the probability of correctly responding to an item when none of the required attributes has been mastered. The two main effects δ_{j1} and $\delta_{j2,}$ show the change in the probability of correctly

responding due to the mastery of attributes α_1 and α_2, respectively. And δ_{j12} shows the interaction effect and hence change in the probability of correctly responding due to mastering both attributes δ_{j1} and δ_{j2}. As stated above, imposing some constraints on the parameterization of G-DINA, namely removing some main or interaction effects from G-DINA or changing its link function, results in the development of specific DCMs.

DINA model

As the simplest interpretable DCM, DINA enjoys a conjunctive/ non-compensatory attribute structure. That is, it requires simultaneous mastery of all the underlying attributes of any given item to result in incremental probability. The probability of correctly responding to item j under the DINA framework is

$$P(X_j = 1 | \alpha_1, \alpha_2) = g_j^{1-\alpha 1\alpha 2}(1- S_j)^{\alpha 1\alpha 2}$$

Where S_j is the probability of a slip, namely, an incorrect response to item j, despite having mastered all the required underlying attributes for the item, and g_j is the probability of a guess, namely, a correct response to item j, despite not having mastered all the required attributes for that item. Setting all the main and lower-order interaction effects to zero results in the development of DINA from the G-DINA model. Hence, the probability of correctly answering an item requiring two attributes α_1 and α_2 for the DINA model can be written as follows:

$$P(X_j = 1 | \alpha_1, \alpha_2) = \delta_{j0}\delta_0 + \delta_{j12}\alpha_1\alpha_2$$

DINO model

In the DINO model, as a disjunctive/ compensatory counterpart to the DINA model, mastery of at least any single attribute increases the probability of correctly answering any given item as mastery of all the required attributes would do. The probability of correctly responding to item j under the DINO framework is

$$P(X_j = 1 | \alpha_1, \alpha_2) = g_j^{(1-\alpha 1)(1-\alpha 2)}(1- S_j)^{1-(1-\alpha 1)(1-\alpha 2)}$$

Similar to the DINA model, $1- S_j$ is the probability of not slipping, and g_j, the probability of guessing for item j. In terms of the parameters in the G-DINA model, it holds that $\delta j_0 = g_j$ and $\delta_j = 1- S_{j0}- S_{j1} = 1- S_{j0}- S_{j1}- \delta_{j2} - \delta_{j12}$.

A-CDM model

Setting all the interaction effects in the G-DINA model to zero results in A-CDM development from the G-DINA model. The probability of correctly answering an item requiring two attributes α_1 and α_2 for ACDM can be written as follows:

$$3. \quad P(X_j = 1 | \alpha_1, \alpha_2) = \delta j_0 + \delta_{j1}\alpha_1 + \delta_{j2}\alpha_2$$

As shown by de la Torre (2011), under the A-CDM framework, each attribute additively contributes to the increment in the probability of a correct response, and a mastered attribute can compensate for the lack of one attribute.

LLM model

LLM is also called compensatory reparameterized unified model (C-RUM) (Ma & de la Torre, 2018, p.41). Like A-CDM, it is developed from the G-DINA by setting all the interaction effects to zero. However, LLM uses a logit link function. The item response probability for a two-attribute item can be written as follows:

$$4.\quad \text{Logit } P\ (X_j = 1 |\ \alpha_{1,}\ \alpha_2) = \delta j_0 + \delta_{j1}\alpha_1 + \delta_{j2}\alpha_2$$

R-RUM model

As with the A-CDM and LLM, R-RUM is also developed from the G-DINA by setting all the interaction effects to zero. But, unlike the A-CDM and LLM, which use identity and logit link functions, respectively, R-RUM uses a log link function. The item response probability for a two-attribute item for R-RUM can be written as follows:

$$5.\quad \text{Log } P\ (X_j = 1 |\ \alpha_{1,}\ \alpha_2) = \delta j_0 + \delta_{j1}\alpha_1 + \delta_{j2}\alpha_2$$

A crucial factor in implementing any of these model-based assessments, which can enhance the validity of inferences made of test data, specifies the cognitive processes variably called abilities, skills, sub-skills, knowledge structure, or attributes that underlie test performance (Kim, 2015). That is attribute definition and attribute specification in a Q-matrix. Despite Afflerbach, Pearson, and Paris's (2008) proposal for conceptualizing the differences between reading skills and strategies by considering two factors of automaticity and intentionality, the authors in this study opted for the conceptualization of the terms provided by the community to merely be consistent with the discourse of the community of cognitive diagnostic assessment. Therefore, mastery in reading comprehension skill, for instance, taking reading as a cognitive domain, requires knowledge of vocabulary, grammar, and making inferences, which are considered the sub-skills of reading domain. The subskills are also called attributes and are used interchangeably throughout the paper.

2.3. Q-Matrix construction and validation

Contrary to developing an assessment tool in which attributes are specified a priori, CDA retrofitting studies require attribute specification from already developed items. Due to the scarcity of cognitive theories underlying test performance in educational assessments, researchers should construct the implicit theory, which can be done in a number of ways. Previous studies have made use of one or a combination of test specifications, theories of a content domain, exploratory approach of item content analysis, introspective or retrospective think-aloud verbal protocols, natural language processing, digital eye-tracking, and brainstorming about possible underlying attributes through test content analysis, and the previously-carried out DCM studies on the construct under study (e.g., Buck & Tatsuoka, 1998; Gorin, 2009; Leighton & Gierl, 2007; Leighton, Gierl, & Hunka, 2004; Lee & Sawaki, 2009a; Ravand, 2016). Still, others have focused on statistical and/or analytical techniques, focusing on surface test task characteristics such as item difficulty to determine the underlying attributes (Sawaki, Kim, Gentile, 2009). Some researchers (e. g., Alderson & Lukmani, 1989) focused "on the hierarchical relationships among L2 receptive subskills in

terms of difficulty to examine whether the existing or suspected hierarchy of attributes could be empirically validated by comparing the hierarchy with actual student performance" (Yi, 2017, p. 2). Still, others have relied on investigating performance differences on the subskills among learners of different levels of proficiency (MeCartty, 1998).

Several issues, however, need to be considered when specifying and defining attributes, including "(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" (Ravand & Baghaei, 2020, p.15). The Q-matrix can be subjectively specified through qualitative analysis using expert judgments (e.g., Jang, 2009; Lee & Sawaki, 2009$_{a;}$ Li, 2011; Ravand, 2016), a practice which was criticized by Gorin (2009). Some researchers have attempted to carry out qualitative analysis in tandem with empirical validation through one of the available empirical Q-matrix validation procedures to address this concern (e.g., Barnes, 2010; Chen, Liu, Xu, & Ying, 2015; Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2016; De Carlo, 2012; Desmarais & Naceur, 2013; Liu, Xu, & Ying, 2012; Templin & Henson, 2006).

In this study, the G-DINA model was selected. Since the results of studies on the nature of reading comprehension subskills relationships are inconclusive, the use of a general, saturated model with its flexible parameterization, which allows for accommodating different kinds of attribute relationships, is warranted. In turn, this feature allows the subsumed models of the general model to compete in being adopted by individual items.

Therefore, the following research questions will be addressed in this study:

1. What attributes/sub-skills are necessary for successfully completing the UEE M.A. reading comprehension (RC) test items?
2. What information will the application of the fitted DCM model to UEE M.A.RC test items provide as to the interaction of attributes within and across items?

3. Method

3.1. Participants and Setting

To obtain and audiotape retrospective verbal reports data by one of the authors and two applied linguistics Ph.D. candidates who were familiar with the methodology, 13 subjects as representatives of University Entrance Examination (UEE) were selected based on purposive sampling and given some monetary incentive to attend an approximately twenty-minute retrospective verbal report one-on-one sessions immediately after completing the reading section of the noted test within 45 minutes. The rationale for their selection was their attendance in 2016 Master of Arts (M.A) examination as ectual test takers because "it is essential that the attribute definition is grounded on empirical investigations of thinking processes (or cognitive operations) underlying the skills and knowledge test takers use to solve educational tasks" (Lee & Sawaki, 2009a, p.176). The exact number of test taker participants for the retrospective verbal report phase was contingent upon data saturation. The authors have opted for this type of data gathering procedure for two reasons: first, it was not logistically feasible to ask the examinees to verbalize their thought processes while completing the test tasks during the high-stakes test administration process. Second,

conducting concurrent think aloud could easily distract examinees during problem solving activity (Ericsson & Simon, 1984) which might, in turn, hinder examinees efficient performance on the test. Moreover, from among all the 17375 examinees taking the test in 2016, responses of approximately 15.11% of examinees (N=2625) were selected for the study. The total score for this sample ranged between 6 and 18 with a mean of 7.77 and a standard deviation of 2.04. Also, four domain experts in two separate panels were involved in coding and rating the underlying attributes. One of the experts was a native English-speaking professor in education[1] and the other three were EFL reading instructors familiar with CDA.

3.2. Instrumentation

Data for this study was collected through a high-stakes national UEE that consists of a specialized content module and a general proficiency module, part of which taps reading comprehension. Access to test takers' answer sheets was possible thanks to the cooperation of the Iranian Measurement Organization (IMO). The test items served two functions. First, they constituted the main elicitation procedure to gather test takers' performance data in the actual UEE M.A. reading comprehension (RC) test administered to applicants holding bachelor's degree and seeking to pursue their studies in one of the English language master's programs namely Teaching English as a Foreign Language (TEFEL) in state universities in Iran held in 2016. The English proficiency test is a multiple choice speeded test that should be completed within 60 minutes. The test typically comprises grammar (10 items), cloze (10 items), vocabulary (20 items) and reading comprehension (20 items). It was the latter part of the test, reading comprehension, which is the focus of this study. The reading tests consist of three expository reading passages for a general adult audience, followed by 20 four-option multiple choice items.

Further data came from a retrospective verbal report procedure, and domain-expert judgment. The same reading items, noted above, were used to elicit retrospective verbal reports from the 13 participants in the retrospective verbal report phase of the study.

As to the reading passages, the first one was a five-paragraph essay of 461 words introducing two research studies done in the area of psychology, working especially on self-esteem and life-satisfaction in the first paragraph. In the subsequent paragraphs, the problems restricting the generalization of the results of the noted studies have been mentioned. Except for the first paragraph, in which the main idea was presented in the last sentence, main ideas for the rest of paragraphs were mostly found at the beginning of each paragraph. The passage was tightly structured and followed by seven questions. The rhetorical structure of the passage was similar to problem-solution according to Grimes' (1975) rhetorical organizer. The second passage consisted of four paragraphs of 417 words. It was about autistic children, misleading image of autistic children, their needs and the ways they can be treated and helped. It was followed by seven questions. This passage can also be classified as 'problem/solution'. The third passage was a five-paragraph essay of 426 words discussing the efficacy of intelligence tests to measure the construct they purport to measure. Main ideas were mostly presented at the beginning of each paragraph except the first paragraph in which the main idea was presented in the last sentence within the paragraph. This passage was followed by 6 questions. The rhetorical organization of this passage was cause-effect.

3.3. Procedures of Q-matrix development

Q-matrix specification, design, and grain size of the attributes (Madison & Bradshaw, 2015; Ravand and Baghaei, 2020) play significant roles in the "classification accuracy, parameter recovery of latent class distributions, correlations, and attribute proportions" (Lei & Li, 2016). To ensure the valid identification of subskills the following steps were taken: (1) retrospective verbal reports of 13 UEE M.A. subjects were gathered and analyzed, (2) two content domain experts, one a native English-speaking professor in education[1] and the other, an EFL reading instructor familiar with CDA, participated in the study to analyze retrospective verbal reports, identify and code the attributes underlying the test items (3) the second panel of content-expert raters, who were also familiar with CDA, judged and rated the underlying attributes, (4) the Q-matrix was empirically validated, and (5) the Q-matrix revision giving the suggestions made by the software package and the options of the expert judges.

Thinking aloud is a challenging practice per se for subjects and, as pointed out by Afflerbach, depends on subjects' ability to verbalize their thinking. Also, asking subjects to think aloud in a foreign language can be even more demanding which can, in turn, hamper the production of rich data. Therefore, since, due to insufficient English speaking potential among examinees, making them to report in English might result in a situation where examinees' verbalization might not match their actual thought processes which, in turn, impeded the production rates of verbal comments and reaching rich data (Afflerbach, 2000; Gass & Mackey; 2000), subjects were allowed to opt for whatever language, i.e. English or Persian, they were comfortable with or to code switch as they choose. The audiotaped reports were later transcribed and translated by one of the authors for further analyses. Then, the first panel of raters provided an estimated task analysis, analyzed retrospective verbal report data, and coded the items independently. Detailed explanation of this stage is provided bellow.

To provide the initial list of attributes, the researchers went through the following process applying Pressley and Afflerbach's (2012) model of 'constructively8responsive reading' and McNamara, Ozuru, Best, O'Reilly's (2007) '4-pronged reading strategies framework'. The rationale for using the noted models had to do with their being both theoretically and empirically well grounded. The first panel created a commonality across all items. For each item, one may assume that prior to answering test questions test takers are involved in the following: setting goals, constructing meaning, identifying main ideas and pertinent details, inferring, visualizing, accessing and using prior knowledge, determining vocabulary meaning, and applying decoding strategies and skills when needed. As this was a test environment, one could assume that test takers would intermittently use strategies for eliminating incorrect choices. We also assumed that test takers accurately construct meaning for the test question and options, as these were texts in and of themselves. There are different forms of re-reading—full on re-reading, scanning, and skimming. Raters' notes of verbal reports indicated re-reading because it was not possible to consistently categorize re-reading as scanning, for example. One could also assume that most readers were drawing on their metacognition, as they set goals, vary the rate of reading in relation to the task at hand, re-read, parse text into manageable chunks, coordinate question-answering routines, check on

suitability of response, and verify their answers. Following this, the first two raters provided the likely item specifics of which two are represented in Table 1 due to space constraints.

Table 1.
Item Specifics for Items 8 and 20 Provided by the First Panel of Raters

Item	Item Specifics
8	Read and comprehend test item; comprehend text (build situation model of text); recall and reference situation model from memory (i.e., what has been comprehended from text); synthesize information from text; choose main idea statement OR Re-read passage and synthesize information; construct main idea statement; match correct answer with main idea AND Possibly conduct metacognitive check to determine that all possible combinations of answer are considered
20	Read and comprehend test item; comprehend text (build situation model of text); recall and reference situation model from memory (i.e., what has been comprehended from text); infer author attitude from constructed meaning OR Re-read paragraph with focus on inferring author attitude AND Possibly conduct metacognitive check to determine that all possible combinations of answer are considered

Stressing the need "to understand how contextual variables influence the availability of information to report and the process of reporting", Afflerbach (2000) provides

> representative aspects of the verbal report methodology that demand comprehensive description, including the characteristics of subjects, texts, tasks, directions to subjects, the transcription of the verbal protocols, the selection of protocol excerpts and their representativeness, the categories used to score think-alouds, and the reliability of coding protocol contents. (p. 171)

In the context of this study, retrospective verbal reports and experts' opinions were complementary in the analysis of the test-takers' cognitive processing in reading. The raters took into account the verbal report data, the stems and options, and the text to code the items. The verbal report data was used along with the first two raters' coding to build the categories and then when they had a single word, i.e., two verbalizations that really communicated the strategy that exemplified the category. After the first panels' independent extraction and coding of participants' retrospective verbal report processes and identification of required subskills, they held several other joint sessions to resolve the likely disagreements. For instance, the attribute 'recall and reference situation model from memory' (i.e., what has been comprehended from text) was a problematic one. Because some students/ examinees who are better readers, they recall and reference what they have read. Others will need to go back. Therefore, better readers will finish reading the text and they have really strong understanding of what it said and then they will answer the question based on the strong understanding. In that case, there will be 'recall and referencing situation model'. If it is a weaker reader, they all go back as directed by the question and work on a situation. Or, not,

they just focus on what the question is. In such cases, the coders decided not to put the attribute in the Q-matrix for two reasons: first, it was not possible to consistently categorize this attribute. Second, the increase in the number of identified attributes while the test length remained the same might have resulted in issues related to identifiability of the DCMs and estimation errors.

As such, they underwent the lengthy qualitative process of discussion sessions on not only the identification of the required sub-skills but also the analysis of the whole processes test-takers went through. It is worth pointing out that in this study the raters came up with the names and categories which seem to be complementary and distinct enough to merit their own category. But, one would challenge us that 'is not all this about test-taking and test-wiseness because it is in a test environment? It is a behavior that otherwise would not be happened?' In fact, there is a rich reading strategy literature based on eye-movement, self-reporting, and verbal reports from readers and they describe the strategies that are fairly common across good readers. However, in this study, the researchers are looking at a special type of reading, i.e., reading a 4-5 paragraph text in a test and everyone in the world can relate to it. But they were going to be focusing on reading strategies while taking a test. So they got to a list out of which the ones which seemed to be more purely reading-oriented along with the experts' opinions were inputted in the initial Q-matrix for DCM application. As such, their analyses of the data yielded two macro-level categories of test-focused and text-focused categories/processes of which only the latter which dealt with processing textual information to solve the tasks were considered for codings to be later inputted in the initial Q-matrix. To further illustrate the analytic procedure, an example of the text-focused attributes is presented. *Example* Building a situation model of the text: referring to the ability to actively process and integrate concepts from the text and related concepts of domain and general world knowledge to construct a coherent mental representation of the content of the text. The following retrospective verbal report excerpts indicate this attribute.

Table 2.
*Ex. erpts of participants' verbal report*s

C. Based on the movie, I mean, my background knowledge, I could remember the story. But to ensure that my answer is correct I scanned again.
E: I reviewed the movie I could remember the scenes…so those 3
H. Question 9…Which of the following is True about Raymond? So if I'm not wrong Raymond is the autistic brother, …. So…ummm…glancing over the options, I believed the third option is correct because he did leave a positive impact on his brother. it's mentioned here that through various experiences it becomes possible for the brother to learn from Raymond and to forge an emotional bond with him. So he did have a positive impact on his brother. So option 3 is correct for question number 9.
M. his T passage was about the 'Raymond' movie. It was about two brothers. One of them suffered from autism…of course, it was not that much acute. After their fathers' death, he inherited his properties …the other brother wasn't good and wanted to deprive him from bequest…but, at the end of the movie, I felt he's liking his brother. So, here, I chose positive impact …that he left a good impression/image of himself.

In the above-mentioned retrospective verbal report excerpts (Table 2), participants C, H, E, and M tried to build a mental image of the character in the passage by recalling the movie they had watched to solve the task. In fact, their inferential comprehension resulted from their attempt to visualize the situations, namely the character and context of the movie, depicted in the text through relating their background knowledge of the movie to the passage helped them find the answer.

Therefore, since both participants' verbalizations and the first panel's independent codings unanimously provided evidence as to the existence of the attribute 'building a situation model of a text', this attribute was put into the Q-matrix. In cases where the participants' and coders' attribute identifications did not match, those of participants were given priority. However, this was applicable only if participants' verbalizations provided the purely reading-oriented subskills. Otherwise, the panel's codings were preferred.

Then during the final joint session, the first panel of coders met to finalize the identification and coding of subskills based on the participants' verbal reports and experts' codings and purely-reading-oriented attributes developed the initial Q-matrix. In this session, they discussed each item again and specified the initial Q-matrix for software application as shown in Table 3.

Table 3.
Q-matrix developed by the first panel of two raters

Item	VOCAB	INF	BSM	TSK	ICM	IRT	SI
1	1	1	1	1	0	1	0
2	1	1	1	0	0	1	1
3	1	1	1	0	0	1	1
4	1	1	1	1	0	1	1
5	1	1	1	0	1	1	1
6	1	1	1	0	0	1	0
7	1	1	1	0	0	1	1
8	1	1	1	0	1	1	1
9	1	1	1	0	0	1	1
10	1	1	1	0	0	0	1
11	1	1	1	0	1	1	1
12	1	1	1	0	0	1	0
13	1	1	1	1	0	0	1
14	1	1	0	0	0	1	0
15	1	1	1	0	1	0	1
16	1	1	1	0	0	1	0
17	1	1	1	1	1	0	1
18	1	1	1	0	0	1	1
19	1	1	1	1	0	0	0
20	1	1	1	0	0	0	0

VOCAB: Vocabulary knowledge; INF: Inferencing; BSM: Building situation model of a text; TSK: Text structure knowledge; ICM: Identifying and constructing main idea; IRT: Identifying relevant text material to answer a question; SI: Synthesizing information from text.

Following this, the first round of empirical Q-matrix validation was carried out applying the current Q-matrix measuring a total of seven attributes, response data of 2625 UEE examinees, and GDINA package Version 2.7.4 (Ma & de la Torre, 2018) in R (R Core Team 2019). Moreover, values of LR, AIC, and BIC were consulted for empirical validation purposes. Overall, this stage resulted in the deletion and removal of the attribute 'identifying relevant text material to answer a question' from the Q-matrix because of its test-taking nature. Meanwhile, the second panel of raters was also provided with the passages followed by test items, the list of identified attributes together with their descriptions and instantiations, to independently provide their ratings on the Q-matrix. Then, a consensus Q-matrix out of ratings of both panels of raters was made for the next round of empirical revisions. In this round, values of LR, AIC, and BIC along with mesa plots were consulted. Finally, considering these values and mesa plots, two panel of raters finalized the Q-matrix presented in Table 9. Therefore, Q-matrix construction and validation phase of the study which underwent multiple revisions yielded six underlying attributes namely, Vocabulary Knowledge (VOCAB), Inferencing (INF), Build a Situation Model (BSM), Text Structure Knowledge (TSK), Identifying and Constructing Main idea (ICM), and Synthesizing Information from Text (SIT), as shown and defined in Table 4 below.

Table 4.
Finalized M.A. test instrument Q-matrix attribute definitions and use across all Items

Attribute	Definition	Item	No of items
Vocabulary Knowledge (VOCAB)	The ability to construct meaning of words using prior knowledge (i.e. domain, topic and general world knowledge), linguistic, and contextual clues	1,2,3,4,5,6, 7,9,10,12,1 3,14,17,19	14
Inferencing (INF)	The ability to assume or make a connection (between textual and contextual elements) which is not explicitly stated either automatically by resorting to prior knowledge to fill in information that is not in the text or non-automatically/strategically by information stuffed in the sentences of the text. the ability to assuming or making a connection (between textual and contextual elements)	2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 16, 17, 18, 19, 20	14
Build a Situation Model (BSM)	The ability to actively process and integrate concepts from the text and related concepts of domain and general world knowledge to construct a coherent mental representation of the content of the text.	4, 7, 8, 9, 11, 12, 15, 16, 17, 19, 20	11
Text Structure Knowledge (TSK)	The ability to discern how the text is structured and organized using prior knowledge, syntactic knowledge and knowledge of the relationships between and among sentences and paragraphs to make inferences about The text, to organize content and to build a mental representation of	1, 13, 17, 20	4

 ISSN 2476-5880

	text content		
Identifying and Constructing Main idea (ICM)	The ability to identify and construct the gist of a paragraph, main idea or title of a passage	5, 7, 8, 11, 13, 15, 17, 18, 20	9
Synthesizing Information from Text (SIT)	The ability to constantly make and recycle intentional bridging (connections or associations) inferences that connect back to previous sentences and ideas in order to construct meaning of a text	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 18	14

Furthermore, the Wald statistics (de la Torre & Lee, 2013) for all the specific models for each item was calculated. In cases where the null hypothesis (the specific model holds for an item) was rejected at $p < .05$, the reduced model is rejected. In case of the retention of more than one reduced model and the existence of DINA, DINO, DINA or DINO with the largest p value was retained. Otherwise, other retained reduced models with the largest p values were selected. The rationale for such selections has to do with the statistically least complex reduced DCMs being preferred over other specific DCMs (Rupp & Templin, 2008).

4. Q-Matrix Validation

After the identification of the initial Q-matrix by the first panel, to begin the quantitative analyses, each UEE's respondent's total score was calculated in Excel and then imported into SPSS for the frequency of each total score to be calculated. Then, the sum of frequencies of scores equal to or above six was calculated. This was done to remove examinees missing most items on the test in order to ensure the adequacy of the diagnostic information (Chen & Chen, 2016, p. 222). The response data of these 2625 examinees to 20 items measuring seven attributes were first analyzed in conjunction with the initial Q-matrix represented in Table 3 using the GDINA package Version 2.7.4 (Ma & de la Torre, 2018) in R (R Core Team, 2019).

Twenty multi-attribute items constitute the current Q-matrix measuring a total of seven attributes. Researchers went through the following process to empirically validate the initial Q-matrix identified by the panel of coders. The G-DINA model with saturated attribute distribution was fitted to the data applying monotonic constraints and the *Qval* function. The researchers used the general Q-matrix validation procedure proposed by de la Torre and Chiu (2016). This validation procedure suits the purpose of the study because it is specifically developed to conform with G-DINA and all the specific DCMs derived from it. Six items out of 20 were subject to 13 modifications. This time, the first panel examined the items and attributes againa They decided, for instance, tozmergel 'identifyingerelevantrtextcmaterial to answer a question' with test-focused strategy of 'reading items first, so that the reading of the text is directed at finding answers' and removed it from the Q-matrix. Their rationale for this modification had to do with the test-taking nature of this attribute. Likewise, neither of the two modifications for Item 10 was supported. Because first, it was not common, in judges'

experience, to associate an author's attitude with 'main idea,' and second, 'identify relevant text material to answer the question' is a test-taking reading behavior, not a pure reading behavior.

At the same time, the software-suggested modifications were applied one at a time and the results of the likelihood ratio (LR) tests were consulted. Overall, results of LR tests showed that $\chi 2$ test (Table 5) was not significant (p>.05) for only the first revision for item 10 (mod 3), i.e., insertion of 'identifying and constructing the main idea' (Table 5W shows, the model with the modified Q-matrix (mod 3) fits better than the model with the original Q-matrix (mod 2), as indicated by the non-significant difference ($\chi 2 = 1.54$, df = 8, p >.05) and lower AIC and BIC values.

Table 5.
Likelihood Ratio tests compared

Model	LL	Deviance	AIC	BIC	$\chi 2$	df	*p*-value
mod2	-33074.71	66149.43	67211.43	70329.91			
mod3	-33042.10	66084.21	67146.21	70264.68	1.54	8	.99

LL: log-likelihood value; AIC: Akaike' information criterion; BIC: Bayesian information criterion; $\chi 2$: Likelihood ratio test; df: the degree of freedom

Meanwhile, the second panel of raters consisting of two other domain experts, familiar with CDA, was provided with the passages followed by test items, the list of identified attributes together with their descriptions and instantiations, to independently provide their ratings on the Q-matrix. Then, a consensus Q-matrix (Table 6) out of ratings of both panels of raters was made for the next round of empirical revisions. Then, the UEE M.A. RC test response data of 2625 examinees to 20 items were analyzed in conjunction with the current Q-matrix measuring a total of six attributes represented in Table 6.

Table 6.
Consensus Q-matrix developed by the board of raters

Item	VOCAB	INF	BSM	TSK	ICM	SI
1	1	0	0	1	0	1
2	0	1	0	0	0	1
3	1	1	0	0	0	1
4	1	1	1	0	0	1
5	1	1	0	0	1	1
6	1	1	0	0	0	1
7	1	0	1	0	1	1
8	0	0	1	0	1	1
9	1	1	1	0	0	0
10	1	1	0	0	0	1
11	0	1	1	0	1	1
12	1	1	1	0	0	0
13	1	1	0	1	1	1

 ISSN 2476-5880

14	1	0	0	0	0	1
15	0	0	1	0	1	1
16	0	1	1	0	0	0
17	1	1	1	1	1	0
18	0	1	0	0	1	1
19	1	1	1	0	0	0
20	0	1	1	1	1	0

VOCAB: Vocabulary knowledge; INF: Inferencing; BSM: Building situation model; TSK: Text structure knowledge; ICM: Identifying and constructing the main idea; SI: Synthesizing information from text

First, to empirically validate the Q-matrix, the G-DINA model with saturated attribute distribution was fitted to the data applying monotonic constraints and the *Qval* function. Results showed six modifications for Items 3, 6, 11, and 17. To further examine the plausibility of the proposed modifications based on de la Torre and Chiu's item-specific discrimination index (ς^2) approach (2016), the corresponding mesa plots for Items 3, 6, 11, and 17 were taken into account. The mesa plot is used to visually specify the best q-vector candidates for each item. It is similar to the scree plot in factor analysis, and is a line chart with the x-axis representing q-vectors having the highest proportion of variance accounted for (PVAFs) for different numbers of required attributes and the y-axis offering the corresponding PVAFs (Ma, 2019). The red solid dot indicates the original q-vector. The correct q-vector for the given item is the one on the edge of the mesa (de la Torre & Ma, 2016). In fact, correct q-matrix is the parsimonious one yielding approximately the highest ς^2 rather than the q-vector producing the highest ς^2 (de la Torre & Akbay, 2019). The parsimonious q-vector is the one that approximates the maximum $\varsigma^2$with the fewest attribute specifications. Results showed that only the original q-vectors for Items 3 and 6 had PVAFs less than .95, indicating that they may need further examination. Figure 1 gives the mesa plots for these two items.

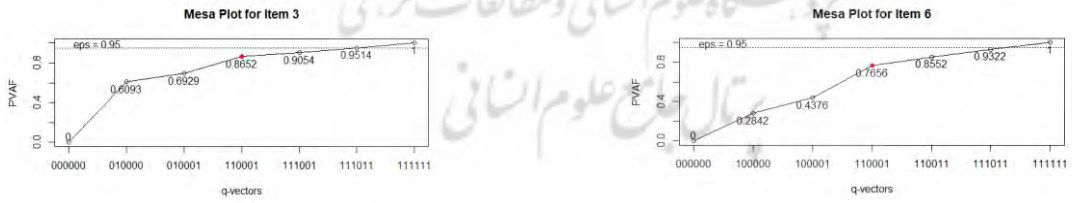

Figure 1. Mesa plots for Items 3 and 6 before applying the revisions

Also, the LR values were consulted to further examine whether the model with the suggested Q (mod 14) had a better relative fit. As Table 7 shows, the model with the modified Q-matrix (mod 14) fits better than the model with the original Q-matrix (mod13), as indicated by the non-significant difference ($\chi 2 = 19e22, df = 12, p > .05$) and lower AIC and BIC values.

 ISSN 2476-5880

Table 7.
Likelihood Ratio tests compared

	LL	Deviance	AIC	BIC	χ2	df	p-value	#Npar
mod 13	-33139.86	66279.72	66877.72	68633.70				299
mod 14	-33149.47	66298.94	66872.94	68558.44	19.22	12	.08	287

LL: loglikelihood value; AIC: Akaike' information criterion; BIC: Bayesian information criterion; χ2: Likelihood ratio test; df: degree of freedom; #Npar: number of parameters

Then, the two panels of raters were also asked for their take on the software-suggested modifications. They seemed modifications for items 3 and 6 plausible. The judges agreed with the revisions suggested regarding Item 3. Initially, they thought that the words *'temperament'* and *'disposition'* would occasion a challenge for the test takers. However, on second thought, they saw test-takers might circumvent the two words. As to the SI attribute, they thought test-takers could get the right answer by just comprehending the section, and no synthesis was required. As to Item 6, they agreed that neither NIF nor SI was required to get the item right. As to Item 11, they did not agree with the suggestion on the grounds that in Paragraph 2, where the answer was located, there was no difficult word. As to Item 17, they did not agree with the suggestion because the test takers needed to understand both paragraphs to see how the two were connected. A decent understanding of both paragraphs requires knowledge of the general and technical words/phrases such as '*conventional, predictive validity, variation, account for…'*

Therefore, the modifications were applied, and empirically examined one by one.

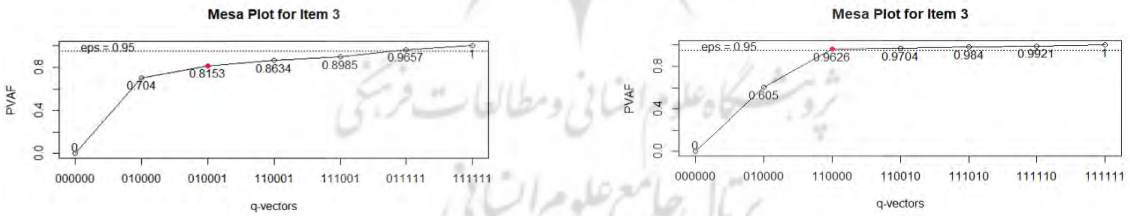

Figure 2. Mesa plots for Item 3 after applying the first and second revisions

After the first revision for item 3, as shown in Table 8, the χ2 test, with 2 degrees of freedom, corresponding to the likelihood ratio tests resulting from comparing the mod15 with the mod16 was significant (p<.05). The result indicates that the mod16 led to a significant loss of fit. Therefore, the first revision for item 3 is not correct. The first revision for item 3 resulted in PVAF value lower than the cutoff. So, as shown in *Figures 1* and *2*, since compared to PVAF value before the revision, the revision resulted in much lower than the cutoff PVAF value; the first revision is not correct. After applying the second revision for item 3 (see Table 8), the χ2 test, with 28 degrees of freedom, corresponding to the likelihood ratio tests resulting from comparing the mod17 with the mod18, was significant (p<.05). The result indicates that mod 18 led to a significant loss of fit. Therefore, the second revision for item 3 is not correct. Moreover, after applying the second revision for item 3 (see *Figure 2*),

 ISSN 2476-5880

the PVAF value exceeds the PVAF cutoff value. Since the *p*-value is larger than zero (<.001), the revision is not correct. As to the first modification for Item 6, the $\chi 2$ test, with 48 degrees of freedom, corresponding to the likelihood ratio test resulting from comparing the mod19 with the mod20, was not significant ($p>.05$). The result indicates that mod 20 fits the data better. Therefore, the modification for Item 6 seems correct. As to the second modification for Item 6, the $\chi 2$ test, with 16 degrees of freedom, corresponding to the likelihood ratio test resulting from comparing the mod21 with the mod22, was significant ($p< .05$). The result indicates that mod 21 fits the data better. Therefore, the second modification for item 6 is not empirically correct.

Table 8.
Likelihood Ratio tests compared

	LL	Deviance	AIC	BIC	$\chi 2$	df	p-value	#Npar
mod 15	-33122.71	66245.42	66835.42	68567.91				295
mod 16	-33159.36	66318.72	66904.72	68625.46	73.3	2	<.001	293
mod 17	-33118.40	66236.80	66826.80	68559.29				295
mod 18	-33171.31	66342.62	66876.62	68444.66	105.81	28	<.001	267
mod 19	-33151.23	66302.46	66892.46	68624.94				295
mod 20	-33170.62	66341.24	66835.24	68285.83	38.78	48	.83	247
mod 21	-33137.12	66274.25	66864.25	68596.73				295
mod 22	-33180.83	66361.66	66919.66	68558.18	87.41	16	<.001	279

LL: log likelihood value; AIC: Akaike' information criterion; BIC: Bayesian information criterion; $\chi 2$: Likelihood ratio test; df: the degree of freedom; #Npar: number of parameters

However, since, as shown in *Figure3*, applying both modifications for item 6 resulted in much lower PVAF values than those of the cutoff values before applying the modifications (see *Figure 2*), neither of the modifications may be acceptable.

 ISSN 2476-5880

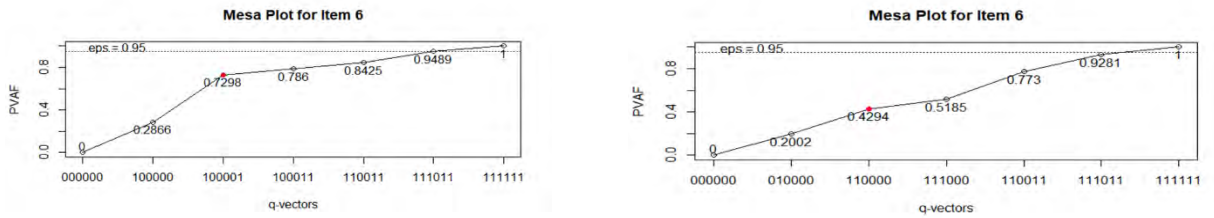

Figure 3. Mesa plot for items 6 after applying the first and second revisions

Therefore, considering the values of LR, mesa plots, and reexamining the items and attributes, the two panels of experts only agreed with the second revision suggested for Item 3 and hence removed attribute SI for this item. Finally, after the noted revisions, the Q-matrix shown in Table 9 was finalized.

Table 9.
Final Q-matrix

Item	VOCAB	INF	BSM	TSK	ICM	SI
1	1	0	0	1	0	1
2	0	1	0	0	0	1
3	1	1	0	0	0	0
4	1	1	1	0	0	1
5	1	1	0	0	1	1
6	1	1	0	0	0	1
7	1	0	1	0	1	1
8	0	0	1	0	1	1
9	1	1	1	0	0	0
10	1	1	0	0	0	1
11	0	1	1	0	1	1
12	1	1	1	0	0	0
13	1	1	0	1	1	1
14	1	0	0	0	0	1
15	0	0	1	0	1	1
16	0	1	1	0	0	0
17	1	1	1	1	1	0
18	0	1	0	0	1	1
19	1	1	1	0	0	0
20	0	1	1	1	1	0

VOCAB: Vocabulary knowledge; INF: Inferencing; BSM: Building situation model; TSK: Text structure knowledge; ICM: Identifying and constructing main idea; SI: Synthesizing information from text

5. Test-level model fit analysis

To examine model fit at the test-level and item-level stages for DCMs, two sets of indices (at each stage), including absolute and relative fit indices, can be consulted (Chen & Chen, 2016; Ma & de la Torre, 2018). A variety of discrepancy-based statistics can assess absolute fit indices at the test-level. Absolute fit measures are evaluated to see whether the model fits the data adequately (see Lei & Li, 2016; Ravand, 2016; Ravand & Robitzsch, 2015, 2018). Regarding max X^2, G-DINA had a non-significant max X^2 value ($p > 0$), indicating a good fit of the model to the data. Considering Maydeu-Olivers' (2013) SRMSR value below .05 as indicating a negligible amount of misfit, G-DINA, as shown in Table 10, fits the data. The MADcor in this study was .0205. DiBello, Roussos, and Stout (2007) considered the MADcor of .049 in Jang (2005) and Roussos, DiBello, Henson, Jang, and Templin (2006); Roussos, DiBello, and Stout (2006) as suggesting a good fit of the DCM to the data. For MADRESIDCOV, MADQ3, values below .05 show a good fit. Except for the MADRESIDCOV value (.45), the value of MADQ3 (.03) was below .05, indicating the fit of G-DINA to the data.

Table 10.
Absolute fit indices for G-DINA

	AIC	BIC	max X^2	p max X^2	MA Dcor	100*MADRESIDCOV	SRMSR	MA DQ3	abs(fcor)	*p*	#Npar
G-DINA	66871.72	68363.42	16.50	.009	.0205	.4580	.025	.0327	.0790	.0049	295

AIC: Akaike' information criterion; BIC: Bayesian information criterion; MADcor: mean absolute difference for the item-pair correlation; MADRESIDCOV: mean residual covariance; SRMSR: standardized root mean score residual; abs(fcor): maximum absolute Fisher-transformed correlation; #Npar: number of parameters

6. Model selection at the item level

Relative fit indices are evaluated to compare rival models for the purpose of selecting the best-fitting model. They are evaluated by using information-based indices such as the Akaike' information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978). The model selection at the item level was carried out to check the fit of the specific DCMs against the G-DINA. As shown in Table 11, after applying monotonic constraints, the results of item-level model fit indicated that the LLM was picked by 13 items, the RRUM by 4 items, the DINO by twoitems, and the DINA by item 12. Compared to the results obtained without applying monotonic constraints, this result also indicates that applying monotonic constraints allows all these 20 multi-attribute items to be selected by a simpler DCM. Following Ravand and Robitzsch (2018), the fit of the multi-DCM model was compared against that of the original G-DINA model, as provided in Table 12, to justify the use of multiple DCM rules for the items within the test. As Table 11 shows, the non-

 ISSN 2476-5880

significant difference (p> .05) and lower AIC and BIC values indicates the multi-DCM fits significantly better than the G-DINA, $\chi 2 = 140.58$, df = 12, p> .05.

Table 11.
Reduced models fitting at item level

Item	DINA	DINO	A-CDM	LLM	R-RUM
1				x	
2				x	
3				x	
4				x	
5				x	
6		x			
7					x
8				x	
9					x
10				x	
11				x	
12	x				
13				x	
14				x	
15				x	
16		x			
17					x
18					x
19				x	
20				x	
Sum	1	2		13	4

Note. x indicates the reduced model fits the item; DINA: *deterministic inputs, noisy "and" gate*; DINO: *deterministic inputs, noisy "or" gate*; A-CDM: additive CDM; LLM: linear logistic model; R-RUM: reduced reparameterized unified model

Table 12.
Likelihood Ratio Test for the G-DINA Model and Nested Specific Models

Model	AIC	BIC	CAIC	SABIC	LL	$\chi 2$	df	p-value	#Npar
Multi-DCM	67104.34	67944.15	68087.15	67489.80	-33409.17	140.58	152	0.74	143
G-DINA	67267.76	69000.24	69295.24	68062.94	-33338.88				295

AIC: Akaike' information criterion; BIC: Bayesian information criterion; LL: log likelihood value; $\chi 2$: Likelihood ratio test; df: the degree of freedom; #Npar: number of parameters

5. Discussion and Conclusion

Contrary to developing a CDA tool in which attributes are specified a priori, CDA retrofitting studies require attribute specification from developed items. The current retrofitting CDA study set out with the aim of illustrating the process of construction and validation of the Q-matrix under the G-DINA model framework. Moreover, the underlying attributes of the high-stakes UEE M.A. RC test items and their interactions were identified and examined. In the same vein, two questions posed in the study: first, what attributes/sub-skills is necessary for successfully completing the UEE M.A. RC test? And second, what information will the application of the fitted DCM model to UEE M.A. RC test items provide as to the interaction of attributes within and across items?

As to the first question, the current study found that six attributes were involved in answering the UEE RC test items; namely, vocabulary knowledge, inferencing, build a situation model of a text, identifying and constructing the main idea, text structure knowledge, and synthesizing information from text. Unlike Rupp, Ferne, and Choi's (2006) study, which explored test-takers' reading behaviors in a multiple-choice RC testing context and non-testing context, this study found the presence of higher order inferences that may lead to an integrated macrostructure situation model in a testing situation. This finding corroborates that of Cohen and Upton's (2007) study in which the think-aloud participants tried to draw on their understanding and interpretation of the passage to answer the questions. However, the results of this study and that of Cohen and Upton's converge with those of Rupp et al. in that the participants in all three studies used test-taking strategies more regularly than those of reading strategies. A detailed look at the models picked by each item and their corresponding attributes seems to provide relevant answers to the second question posed. The results of item-level selection showed that all twenty multi-attribute UEE RC test items variously were held by specific reduced models: The LLM was picked by 13 items (Items 1,2,3,4,5,8,10,11,13,14,15,19, and 20), the R-RUM by four items (Items 7,9,17, and 18), the DINO by two items (Items 6 and 16), and the DINA only by Item 12. The adoption of all the items by a simpler model helps in interpreting the relationships among their attributes and, in the case of correct adoption, results in more accurate classifications (Rojas, de la Torre, & Olea, 2012). Among the 4 two-attribute Items 2, 3, 14, and 16, the first three were picked by LLM, and the last one by DINO, enjoying additive and disjunctive compensatory attribute relations, respectively. The most dominant pair of attributes was VOCAB/INF measured in 10 items and mostly seen in combination with at least one other attribute, except in item 3. Both Items 8 and 15, measuring three attributes of BSM, ICM, and SI, were picked by LLM. Adding attribute INF to this combination in Item 11 did not change the model picked. However, adding VOCAB resulted in Item 7 to be picked by another additive model, namely, R-RUM. The same combination of the attributes VOCAB, INF, and BSM resulted in Items 9, 12, and 19being picked by additive R-RUM, conjunctive non-compensatory DINA, and additive LLM models, respectively. Adding SI to this combination of attributes, however, resulted in Item 4 to be picked by another additive model, namely, LLM. A likely explanation for these findings, as Yi (2017) points out, might be "that the contribution of a particular attribute varies more across items than the contribution of each

attribute varies within an item" (p. 12). A possible explanation for the result that all 4- and 5-attribute items were picked by additive models (RRUM, LLM) might be that as the number of attributes within/ across items increases, the more likely it is that an additive (compensatory) model best fits the corresponding items. Here, the adoption of items measuring more attributes by additive models might be challenged solely due to their having most parameters and not to their compensatory nature. Two pieces of counter-evidence, however, can refute this speculation. The first has to do with the fact that each item, even in the same test, is allowed to be picked by the best-fitting specific model under the G-DINA model framework. In fact, additive models were picked by the majority of items (Table 11). Another piece of counter-evidence is that G-DINA, despite having a saturated structure and, in turn, more parameters than any specific models, was not picked by any item. One possible explanation for this resulting from the obtained values of AIC and BIC may be that overly complex models like G-DINA that produce only a small improvement in fit are penalized, hence not yielding better fit compared to the more simply structured additive models, for instance, LLM.

The selection of the majority of the items by two additive models, namely, LLM (13 Items) and R-RUM (4 Items), which is indicative of the best fit of these models to most items (17 Items out of 20), can be translated into the fact that the processing of UEE's L2 reading skill can best be mirrored by the additive modeling scheme, specifically LLM. Table 13 helps contextualize these findings by showing different items measuring different combinations of attributes being variously picked by specific models. As to the nature of processing and interaction of the specified attributes, the fact that the LLM was picked by most items (N=13) suggests that UEE L2 reading attributes favor a compensatory relation among the specified attributes. This finding corroborates the finding of Ravand and Robitzsch's CDA study (2018) on UEE RC test items.

Table 13.
Item-specific models defined by the G-DINA for items measuring different combinations of attributes

Model	Type	Link function	
LLM	Additive Main effects	logit	Item 1(VOCAB,TSK,SI), Item 2 (INF, SI), Item 3 (VOCAB,INF), Item 4(VOCAB, INF, BSM, SI), Item 5 (VOCAB, INF, ICM, SI),item 8 (BSM, ICM, SI), item 10 (VOCAB,INF, SI), item 11(INF, BSM, ICM, SI), item 13(VOCAB, INF, TSK,ICM, SI), item 14(VOCAB, SI), item 15 (BSM, ICM, SI), item 19 (VOCAB, INF, BSM, TSK, ICM), item 20 (INF, BSM, TSK, ICM)
R-RUM	Additive Main effects	log	Item 7 (VOCAB, BSM, ICM, SI), Item9 (VOCAB, INF, BSM), item17 (VOCAB, INF, BSM, TSK, ICM), item18 (INF, ICM, SI)
DINO	Compensatory parsimonious	Identity	Item 6 (VOCAB, INF, SI), item 16 (INF, BSM)
DINA	Non-compensatory parsimonious	Identity	Item 12 (VOCAB, INF, BSM)
A-CDM	Additive Main effects	Identity	-

However, a number of limitations need to be considered. First, the main weakness of this study, as with most other retrofitting CDA studies, deals with the use of a non-cognitive-diagnostic test without subskill specification for DCM application. Although the authors tried their utmost to provide an ecologically valid and authentic Q-matrix through triangulation of the data using different sources, using a non-cognitive-diagnostic test with subskill specification as an alternative for retrofitting study can be suggested. The second limitation has to do with the design of the attributes in the specified Q-matrix. Madison and Bradshaw (2015) argue that the correct specification of the Q-matrix is vital but not sufficient for classification accuracy. Another equally important factor that may influence classification accuracy is Q-matrix design. For instance, in this study, attributes VOCAB/ INF were always measured together, making classification accuracy suffer. Moreover, none of the attributes were measured in isolation, which may otherwise increase classification accuracy. As a solution to these problems, one may examine the effect of various Q-matrix designs to improve classification accuracy. As an alternative solution, Madison and Bradshaw recommend that each attribute be measured with other attributes in case it cannot be measured in isolation (as in VOCAB INF 000 in our case). A third solution offered is merging the two attributes to form a composite attribute in case "two attributes are truly attached, and items cannot be written to measure either attribute without the other" (Madison & Bradshaw, 2015, p. 509). However, in such cases, more caution must be exercised so as not to violate the substantive considerations at the cost of reaching solely favorable statistical values since, for instance, two attributes may substantively be distinct enough to merit their own categories despite their always coming in conjunction. Put it differently, the application of such recommendations is justifiable only if they are not counter to domain-specific theoretical considerations. Finally, more complex Q-matrices, i.e., Q-matrices with most of the entries filled by 1, may jeopardize estimation capabilities culminating in reduced effectiveness of the model to accurately classify respondents. Taken together, the practice of triangulation of data and anticipation of the likely noted problems in the process of developing and validating the Q-matrix, especially at the design stage of CDA assessment tool development, will hopefully contribute to more substantively valid inferences.

Acknowledgments

The authors would like to thank Professor Peter Afflerbach for his collaboration in the analysis of the verbal report data and coding the attributes.

Declaration of Conflicting Interests

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

References

Afflerbach, P. (2017). *Understanding and Using Reading Assessment, K–12, 3rd Edition*: ASCD.

Afflerbach, P. (2016). How the tests used in evaluating reading misrepresent student development and teacher effectiveness. In R.E. Gabriel & R.L. Allington (Eds.),

Evaluating literacy instruction: Principles and promising practices (pp. 31–43). New York, NY: Guilford.

Afflerbach, P. (2004). *High stakes testing and reading assessment*. Commissioned Policy Brief for the National Reading Conference.

Afflerbach, P. (2000). Verbal reports and protocol analysis. In (Eds.) M. L. Kamil, P.B. Mosenthal, P.D. Pearson, R. Barr, *Handbook of Reading Research* (Vol. Volume 3, pp. 163-179): Taylor & Francis

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher, 61*(5), 364-373. doi:10.1598/RT.61.5.1

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. doi:10.1109/TAC.1974.1100705

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

Barnes, T. (2010). Novel derivation and application of skill matrices: The q-matrix method. In C. Romero, S. Ventura, M. Pechenizkiy, R.S.J. Baker (Eds.), *Handbook of Educational Data Mining* (Vol. Volume 3, pp. 159-172): Taylor & Francis

Barnett, M. A. (1989). *More than meets the eye: Foreign language reading. Language and Education: Theory and Practice*: ERIC.

Bernhardt, E. B. (1986). "Proficient tasks or proficient readers?". *ADFL Bulletin, 18(1)*, 25-28.

Bernhardt, E. B. (2010). *Understanding advanced second-language reading*: Taylor & Francis.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing, 15*(2), 119-157. doi:10.1177/026553229801500201

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment* (Vol. 184): SAGE Publications, Incorporated.

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

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*(510), 850-866. doi:10.1080/01621459.2014.934827

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598-618. doi:10.1177/0146621613488436

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

Cohen, A. D., & Upton, T. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®. *Language Testing, 24*(2), 209-250.

Dana R, F., & Hedgcock, J. S. (2009). *Teaching readers of English: students, texts, and contexts.*NewYork:Routledge.http://125.234.102.146:8080/dspace/handle/DNULIB_52011/3172

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. doi:https://doi.org/10.1111/j.1745-3984.2008.00069.x

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

de la Torre, J. & Akbay, L. (2019). Implementation of cognitive diagnosis modeling using the GDINA R package. *Eurasian Journal of Educational Research. 80*(2019), 171-192. Doi:10.14689/ejer.2019.80.9

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

de la Torre, J., & 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*(4), 355-373. doi:https://doi.org/10.1111/jedm.12022

de la Torre & Ma, (2016). Cognitive diagnosis modeling: A general framework approach and its implementation in R. In M. von Davier, Y.-S. Lee (Eds.), *Handbook of diagnostic classification model, Methodology of Educational Measurement and Assessment*,(pp. 593-601): Springer Nature Switzerland AG. http://doi.org/10.100/978-3-030-05584-4_29

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*(2), 89-97. doi:https://doi.org/10.1016/j.pse.2014.11.001

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *36*(6), 447-468. doi:10.1177/0146621612449069

Desmarais, M. C., & Naceur, R. (2013). *A matrix factorization method for mapping items to skills and for enhancing expert-based Q-matrices.* Paper presented at the Artificial Intelligence in Education, Berlin, Heidelberg.

Dibello, L. V., Roussos, L., & Stout, W. F. (2007). Review of cognitively diagnostic Assessmentand a summary of psychometric models. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics. Psychometrics,* (Vol. Volume 26, pp. 979-1030): Amsterdam: North-Holland Publications

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis.* Cambridge, MA: MIT Press/ Bradford

Ghaith, G. (2018). Reading comprehension instructional framework. *TESL Reporter, 52*(2), 1-17.

Gorin, J. S. (2009). Diagnostic classification models: Are they necessary? Commentary on Rupp and Templin.*Measurement: Interdisciplinary Research and Perspectives, 7*(1), 30-33. doi:10.1080/15366360802715387

Grabe, W. & Jiang, X. (2013). Assessing reading. In A. J. Kunnan (Ed), *the companion to language assessment*, (pp. 1-16): John Wiley &Sons, Inc. doi: 10.1002/ 9781118411360.wbcla200

Grabe, W. & Stoller, F. L. (2013). *Teaching and Researching Reading (2nd Ed.).* New York: Rutledge

Grimes, J. E. (1975). *The Thread of Discourse*: De Gruyter Mouton.

Hartz, S. (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., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*(4), 262-277. doi:10.1177/0146621604272623

Hemati, S. J., & Baghaei Moghadam, P. (2020). A Cognitive diagnostic modeling analysis of the English reading comprehension section of the Iranian national university entrance examination. *International Journal of Language Testing*. *10*(1), 11-32. Retrieved from http://www.ijlt.ir/article_114278_022ab6b64d4bc8acba9c561846828270.pdf

Hemmat, S. J., Baghaei, P., & Bemani, M. (2016). Cognitive diagnostic modeling of L2 reading comprehension Ability: Providing Feedback on the Reading Performance of Iranian Candidates for the university entrance examination. *International Journal of Language Testing*. *6*(2), 92-100. Retrieved from http://www.ijlt.ir/article_114432_4a06dd89576d5c256bd91dea7a5cb398.pdf

Henson, R., & Douglas, J. (2005). Test Construction for Cognitive Diagnosis. *Applied Psychological Measurement, 29*(4), 262-277. doi:10.1177/0146621604272623

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191. doi:10.1007/s11336-008-9089-5

International Literacy Association. (2017). *The role of standardized reading tests in schools* [Literacy leadership brief]. Newark, DE: Author.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL* (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), 031-073. doi:10.1177/0265532208097336

Javidanmehr, Z., & Anani Sarab, M. R. (2017). Cognitive diagnostic assessment: Issues and considerations. *International Journal of Language Testing*. *7*(2), 73-98. Retrieved from http://www.ijlt.ir/article_114441_0f0fe9a2917419485010d66e7db8921c.pdf

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

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. doi:10.1177/0265532214558457

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*(3), 172-189. doi:10.1080/15434300902985108

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

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

Leighton, J., Gierl, M., & Hunka, S. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*. doi:10.1111/j.1745-3984.2004.tb01163.x

Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow, 9*, 17-46.

Li, H., Hunter, C. V., & Lei, P.-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing.33*(3), 391-409. doi:10.1177/0265532215590848

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

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), 357-383. doi:10.1177/0013164416685599

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*(7), 548-564.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Rutledge.

Ma, W. (2019). A diagnostic tree model for polytomous responses with multiple strategies. *British Journal of Mathematical and Statistical Psychology, 72*(1), 61-82.

Ma, W., & de la Torre, J. (2018). GDINA: The generalized DINA model framework.R package version 3.5.1. Retrieved from https://CRAN.R-project.org/package=GDINA

Madison, M., & Bradshaw, L. (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. doi:10.1177/0013164414539162

Magliano, J. P., Millis, K. K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate assessment tools. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies, (pp. 107-136). Mahwah, NJ*: Lawrence Erlbaum Associates

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

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

McNamara, D. S., Ozuru, Y., Best, R., O'Reilly, T. (2007). The 4-pronged comprehension strategy framework. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies,(pp. 465-496). Mahwah, NJ*: Lawrence Erlbaum Associates

MeCartty, F. H. (1998). The effects of proficiency level and passage content on reading skills assessment1. *31*(4), 517-534. doi:https://doi.org/10.1111/j.1944-9720.1998.tb00597.x

National Assessment Governing Board. (2015). *Reading framework for the 2015 National Assessment of Educational Progres*. Washington, DC: National Assessment Governing board, US Department of education. Retrieved from http://www.nagb.org/publications/frameworks/reading/2015-reading-framework.html

Organization for Economic Co-operation and Development. (2014). *PISA 2012 results: What students know and can do—student performance in mathematics, reading, and science*. OECD Publishing. Retrieved from http:// dx.doi.org/10.1787/9789264208780-en

Pressley, M., & Afflerbach, P. (2012). *Verbal protocols of reading: The nature of constructively responsive reading*: Taylor & Francis.

R Core Team (2019). R: A language and environment for statistical computing. R foundation for statistical computing, Viena, Austria. Retrieved from http://www.R-project.org/

Rajagopalan, K., & Gordon, E. W. (2016). *The testing and learning revolution: The future of assessment in education*: Springer.

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

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

Ravand, H., Barati, H., & Widhiarso, W. (2013). Exploring diagnostic capacity of a high stakes reading comprehension test: A pedagogical demonstration. *International Journal of Language Testing. 3*(1), 11-37. Retrieved from http://www.ijlt.ir/article_114382_5823ffa0c6659070c356c3e069838822.pdf

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

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

Rojas, G., de la Torre, J., & Olea, J. (2012). *Choosing between general and specific cognitive diagnosis models when the sample size is small.* Paper presented at the annual meeting of the National Council of Measurement in Education, Vancouver, British Columbia, Canada.

Roussos, L. A., DiBello, L. V., & Stout, W. (2006). Diagnostic skills-based testing using the Fusion-Model-based Arpeggio system. In J. Leighton & M. Gierl (Eds.), *Cognitive*

diagnostic assessment for education: Theory and applications: Cambridge University Press. Doi: 10. 1017/CBO9780511611186.010

Roussos, L. A., DiBello, L. V., Henson, R. A., Jang, E. E., & Templin J. L., (2006); Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. Embretson & J. Roberts (Eds.), *New directions in psychological measurement with model-based approaches* (pp. 35-69). Washington, DC: American Psychological Association.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441-474.

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

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

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. doi:10.1080/15434300902801917

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 1-17. doi:10.1080/15366367.2018.1435104

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

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

von Davier, M. (2005). A general diagnostic model applied to language testing data (ETS Research Report no. RR-05-16). Princeton, NJ: Educational Testing Service.

Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item-level fit for the DINA model. *Applied Psychological Measurement, 39*(7), 525-538. doi:10.1177/0146621615583050

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

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