Journal of English Language
Teaching and Learning
University of Tabriz
No. 21, 2018

Constructing and Validating a Q-Matrix for Cognitive Diagnostic Analysis of a Reading Comprehension Test Battery*

Mohammad Alavi**

Associate professor of University of Tehran

Fatemeh Ranjbaran

PhD Candidate, University of Tehran (Corresponding Author)

Abstract

Of paramount importance in the study of cognitive diagnostic assessment (CDA) is the absence of tests developed for small-scale diagnostic purposes. Currently, much of the research carried out has been mainly on large-scale tests, e.g., TOEFL, MELAB, IELTS, etc. Even so, formative language assessment with a focus on informing instruction and engaging in identification of student’s strengths and weaknesses to guide instruction has not been conducted in the Iranian English language learning context. In an attempt to respond to the call for developing diagnostic tests, this study explored developing a cognitive diagnostic reading comprehension test for CDA purposes. To achieve this, initially, a list of reading attributes was prepared based on the literature and then the attributes were used to construct 20 reading comprehension items. Then seven content raters were asked to identify the attributes of each item of the test. To obtain quantitative data for Q-matrix construction, the test battery was administered to 1986 students of a General English Language Course at the University of Tehran, Iran. In addition, 13 students were recruited to participate in think-aloud verbal protocols. On the basis of the overall agreement of the content raters’ judgements concerning the choices of attributes and results of think-aloud verbal protocol analysis, a Q-matrix that specified the relationships between test items and target attributes was developed. Finally, to examine the CDA of the test, the Fusion Model, a type of cognitive diagnostic model (CDM), was used for diagnosing the participants' strengths and weaknesses. Results suggest that nine major reading attributes are involved in these reading comprehension test items. The results obtained from such cognitive diagnostic analyses could be beneficial for both teachers and curriculum developers to prepare instructional materials that target specific weaknesses and inform them of the more problematic areas to focus on in class in order to plan for better instruction.

Keywords: Attributes; Cognitive Diagnostic Assessment; Cognitive Diagnostic Model; Fusion model; Think-aloud verbal protocol; Q-Matrix

*Received date: 2016/12/10 Accepted date: 2018/06/01

**E-mail: smalavi@ut.ac.ir

Introduction

Until recently, testing and assessment in general have been employed as a measure to obtain overall, average or individual scores of achievement for examining the tenant of accountability. However, assessment that can function as formative to inform instruction has currently become the focus of attention for diagnostic purposes and to embark on strengths and weaknesses of students to guide instruction. From another stance, teacher-made tests have recently gained attention because of their functions for formative assessment in students' learning (Huang & Wu, 2013). To facilitate the process of learning, teachers are expected to be competent in test construction and learning diagnoses in class. As these tests are expected to detect student errors during the learning process, the use of diagnostic tests to improve student's conceptual understandings has been highly valued and recognized in many fields (Hartmann, 2001). Therefore, researchers and practitioners have focused on combining cognitive psychology and educational measurement to enhance learning and instruction (Leighton, Gierl and Hunka, 2004; Mislevy, 2006; Snow & Lohman, 1989; Tatsuaoka, 1995).

As believed by many scholars, reading is an important skill for gaining knowledge in all fields of the academic context. Thus, it is crucial to examine the different components of reading ability and reading skills in order to better understand this skill and to find the related problems of language learners. By diagnosing problematic areas of a reading skill during the course of a term, required feedback can be provided in order to improve learning and make up for students' deficiencies.

There is much critique of the main goal of educational tests that provide quantitative assessment of a student's general ability and proficiency as compared to other student's in the normative group. This type of norm-referenced testing has been used to a great extent for the ranking and selection of students to make educational decisions. In addition to providing merely general information about student's skills and their ability to perform on a test, these assessments are less capable of providing detailed information about student's strengths and weaknesses that could possibly help them in improving their skills or that might even assist the teacher in instructional planning or serve as a guideline for their teaching. Recently, scholars suggest that it is cognitive diagnostic assessment that has a key role in improving the informational value of assessment (Alderson, 2010; de la Torre, 2009; Jang, 2005; Leighton and Gierl, 2007; Rupp et al., 2012). In his commentary on "*Cognitive Diagnosis and Q-Matrices in Language Assessment,"* Alderson (2010) pinpoints his disappointment of the fact that there was no discussion on there being very few truly diagnostic tests in existence. In fact, nearly all studies carried out to date have been on existing large-scale assessments, and no tests have been constructed in order to carry out cognitive diagnostic analyses. He adheres that far more studies have been invested on developing and researching high-stakes proficiency tests than are devoted to any other type of test, namely placement, achievement, or aptitude, nonetheless those specifically constructed for cognitive diagnosis in the form of classroom-based or formative assessments. This study responds to the call for cognitive diagnostic assessment of a newly devised diagnostic test, one that will attempt to provide detailed information about student's strengths and weaknesses in reading comprehension.

Literature Review

In CDA, the different components of a specific domain (in this case, reading) are referred to as attributes. Attributes are the divided components of a cognitive ability, which can be defined as "procedures, skills, or knowledge a student must possess in order to successfully complete the target task" (Birenbaum, Kelly, & Tatsuoka, 1993, p.443). Therefore, L2 reading attributes are composed of different types of language knowledge, skills, and strategies, which are required in comprehending texts (Templin, 2004; Birenbaum et al., 1993).

In the assessment of reading comprehension in a foreign language, the many underlying cognitive attributes have made it a complex process. Reading ability is an important tool for gaining knowledge and improving learning in everyday academic settings and everyday life in general. Therefore, it comes to no surprise that the nature of reading ability has been the focus of research in applied linguistics, education and psychology for quite some time (Cohen & Upton, 2006). Regardless of the extensive research on reading ability, there is still some debate as to how second language reading ability is defined and how its performance should be analyzed and reported. It seems that teachers, students, and practitioners have not been given diagnostic feedback that could be used for improvements in reading ability. These are issues that mostly need consideration in the context of second language reading assessment.

At times, there is such emphasis on reading strategies, that other important elements such as language knowledge, including pragmatic knowledge and grammatical knowledge, have been ignored. One aspect of second language reading ability is the use of language to understand written text. Therefore, both aspects of language knowledge and strategic competence should be considered in order to understand written texts. While the difficulty of defining the construct of reading ability is clear, other problems have been seen with regards to how L2 reading performance has been analyzed and reported. L2 reading test scores are often reported using a general test score without any detailed information (Goodman & Hambleton, 2004). When exams provide only one total score, it can serve the test’s immediate purpose; however, it cannot be used to improve reading performance (Stiggins, Alter, & Chappius, 2004). Only providing a total score does not provide information regarding each student’s specific strengths and weaknesses (Sheehan & Mislevy, 1990). On the other hand, a detailed score report of each individual, including their performance on each reading component at the item level, can be used to both improve reading ability and guide instruction (Snow & Lohman, 1989).

Frameworks for Developing Cognitive Diagnostic Tests

Embretson’s Cognitive Design System (CDS) (Embretson & Gorin, 2001) and Mislevy’s Evidence-Centered Design (ECD) (Mislevy, 1994; Mislevy, Steinberg, & Almond, 2002) are two of the approaches more generally used for diagnostic test development. Considering the issues of construct definition through item writing, and concluding with validation procedures rather than the CDM statistical models, the two approaches focus on the use of cognition in the process of item and test development (Leighton & Gierl, 2007).

Both approaches may differ in their emphasis on the different parts of assessment design and their details, but both share the three principles of the assessment triangle. The assessment triangle includes three related elements that are cognition, observation and interpretation (Pellegrino, Chudowsky and Glaser, 2001). This panel of researcher's believe that cognition is related to a cognitive model about how students represent knowledge and how they develop competence in a certain subject (p.44). A cognitive model provides a description of what should be assessed, but it is different to some extent. According to Leighton & Gierl (2007), a cognitive model specifies the cognitive components and processes, which constitute the construct being tested. This leads to more detailed specifications that are more applicable for instructional feedback. The final key point is that these specifications are backed up by a cognitive theory, meaning that a model about specific cognitive processes related to the construct being tested empirically supports the model.

Cognitive Diagnostic Models

Cognitive Diagnostic Models (CDMs) are data analysis techniques that are designed to link cognitive theory with the items' psychometric properties (Leighton & Gierl, 2007). Most of the studies carried out thus far have been mostly limited to analysis of existing tests, not the development of new assessments, while this study focuses on a cognitive diagnostic assessment of a test developed based on a cognitive diagnostic framework.

Among CDMs, Tatsuoka's Rule Space Model (Tatsuoka, 1995), the Attribute Hierarchy Method (AHM) (Leighton, Gierl, & Hunka, 2004) and the Fusion Model (Hartz, 2002) can be referred to. Most CDMs are IRT-based latent-class models in which the characteristic of multidimensionality is the most important. In previous uni-dimensional IRT-models, examinee ability was modelled by a single general ability parameter. This is while the multidimensionality trait of CDMs makes it possible to investigate the mental processes underlying the student's response by breaking the overall ability down into different parts. The number of dimensions depends on the number of skill components involved in the assessment. The latent variables of CDMs consist of dichotomous, such as mastery or non-mastery, or polytomous levels, such as a rating variable with values such as excellent, good, fair, poor, etc. The loading structure of a CDM is the Q-matrix, which maps the skills necessary to successfully answer each item on the test (Li & Suen, 2013).

Fusion Model

The Fusion model is a type of cognitive diagnostic model that is used to make inferences about the mastery level of each attribute for each examinee, based on the examinees item responses (Dibello & Stout, 2008). In other words, the fusion model is an IRT multidimensional model, also known as the reparameterized unified model that expresses the stochastic relationship between item responses and underlying skills as follows (DiBello et al., 1995):

$$P(X_{ij} = 1|\bar{\alpha}_j, \theta_j) = \pi_i^* \prod_{k=1}^{k} r_{ik}^{*(1-\alpha_{jk})q_{ik}} p_{ci}(\theta_j)$$

in which, $P(X_{ij} = 1/\alpha_j, \theta_j; \pi_i*, r_{ik}*, c_i)$ is the probability of person *j* on item *i* scoring a correct response (X=l) instead of an incorrect one (X=0), given person abilities— α_j, θ_j —and item parameters π_i*, $r_{ik}*$, c_i.

Developing the Q-Matrix

The initial step in generating diagnostic information to help instruction is to map test items onto an item-by-skill table known as the Q-matrix. It consists of an $i \times k$ matrix of binary information in ones and zeros, where *i* is the number of items and *k* represents the number of attributes. A Q-matrix is a representation of a hypothesis regarding which skills are necessary to answer each item on the test (Li & Suen, 2013). So each item will most likely require more than one skill to be answered correctly. According to Buck et al. (1998) developing a Q-Matrix requires following a certain procedure. First a list of skills is developed and then each item is coded based on what skills are required for each item. The first step of the process is done by referring to the previous literature and referring to content expert's judgment and verbal reports on the underlying skills for each item. The next step is to analyse the data using a Cognitive Diagnostic Model, in this case the Fusion Model, with the developed Q-Matrix. Finally, the Q-Matrix is modified based on statistics for each skill.

In developing a Q-matrix, a number of alternatives exist. One of the less costly and efficient approaches is to use existing test specifications; however, the attributes indicated are usually too general for diagnostic purposes. According to Leighton and Gierl (2007), relying on existing test specifications for Q-matrices is usually unwarranted, so in this study an attempt was made at developing a reading comprehension test based on a cognitive diagnostic framework, with the help of Q-matrix construction. Jang (2009) suggests using data from student's verbal reports to construct the Q-matrix. Even though there are doubts as to the validity of verbal reports, they are considered as fairly reliable and useful for reading research (Leighton and Gierl, 2007). Another approach is to use a group of experts to describe the underlying cognitive skills needed to answer each question, based on their previous experience in this realm (Sawaki, Kim & Gentile, 2009). According to Leighton & Gierl (2007) an underlying problem with this approach is the higher ability level of the experts compared with the students, resulting in a gap between the skills and processes truly used by the students and those proposed. Nevertheless, studies on Q-matrix construction have indicated that using content expert judgment in Q-matrix construction does increase its reliability (Jang, 2005, 2009; Kim, 2015; Li, 2011; Li & Suen, 2013; Svetina, Gorin, & Tatsuoka, 2011; Sawaki et al., 2009).

After developing the initial Q-matrix, large-scale data can be used to empirically validate the Q-matrix based on the initial results of cognitive diagnostic modelling. Often, attributes that are similar are combined to reduce the number of attributes. For example, Jang (2009) refined her initial Q-matrix of the LanguEdge reading comprehension test by reducing the number of entries based on Fusion analysis results.

Another instance is the Q-matrix construction done by Kim (2015). This study followed Hartz (2002), in which attributes that were measured by fewer than three items were either merged with similar attributes or deleted from the Q-matrix.

Research Questions

The purpose of this study is to develop a reading comprehension test based on a cognitive diagnostic framework and validate it by constructing a Q-matrix for the identification of underlying skills necessary to respond to the test items.

1. What reading skills are assessed by the newly developed reading comprehension test and how frequent are they for each item of the test?

2. To what extent can the developed items discriminate masters from non-masters?

3. What are the relationships in the participant's performances on the attributes across different levels of beginner, intermediate and advanced?

Methodology

Participants

1986 students from the University of Tehran took part in the reading comprehension test. They were bachelor's students of various majors taking part in the General English course, a requirement of the bachelor's program at the University of Tehran. Response data from the 1986 examinees to each of the 20 questions of the developed test were used for empirical validation.

Thirteen B.A. students (nine female and four male) were recruited to understand their use of the reading skills through a think-aloud verbal protocol. Each of the students had previously taken the reading comprehension test and was asked to verbalize his/her thought processes when answering the test items. This included what skills and strategies first came to mind when answering the items. The reading passages and items were given as reference.

A group of content raters served as developers of the attributes that reflect the main language skills necessary for successful performance on each item. This group included 6 PhD students at the University of Tehran studying Teaching English as a Foreign Language, 3 females and 3 males, and who had experience in applied linguistics research and teaching reading comprehension courses. They reviewed each test item and selected attributes provided and decided whether or not each attribute was necessary for answering the item correctly. They examined the extent to which the attributes specified from different sources, including the verbal reports, are distinguishable from each other and whether they agree upon the attributes associated with the test item. Suggestions from this group were used to develop the Q-matrix, alongside the previous literature in this realm and think-aloud verbal reports.

The test takers (N=1986) test performance data were used to understand item characteristics and then to refine the Q-matrix by the use of statistical modelling through the Fusion model.

L2 Reading Attributes

An issue of great significance in attribute specification is the number of attributes defined for items on a specific test. CDA specifies attributes in a fine-grain size because these skills enhance the cognitive processes underlying the test. According to Rupp et al. (2012), the grain-size of an attribute is the level of precision that one intends to use in analyzing a cognitive response process and report on its constituent components. Coarse-grained descriptions of attributes and cognitive processes are often used in tables of specifications or blueprints for educational assessments (Rupp et al., 2012). On the other hand, fine-grained attributes are used in standards-based assessments with the aim of connecting evidence of learning to the outcome of learning to provide feedback for instruction (Leighton & Gierl, 2007). The objective of the diagnostic assessment and the level of precision that we would like to make assertions about test takers determine the adequate grain-size. There are no concrete standards for the number of attribute labels for a specific test. Even though CDMs can measure an infinite number of attributes, in a practical sense an upper limit of 10 attributes is appropriate, due to the number of possible combinations of items possible (Roussos et al., 2007). It has been suggested that every attribute should be assessed by at least three items, making the results much more interpretable. The list of reading attributes in this study (as shown in Table 1) was developed based on previous literature (Jang, 2009; Cohen & Upton, 2006; Francis et al., 2006; Birch, 2002; Fletcher, 2006; Rupp et al., 2006), content expert judgment, and examinees' think-aloud verbal protocols. The list included those L2 reading attributes that were considered to be involved in the reading process.

Table 1. Attributes of L2 reading ability

	L2 Reading Attributes
Attribute 1	determining word meaning from context
Attribute 2	determining word meaning out of context
Attribute 3	comprehending text-explicit info
Attribute 4	comprehending text-implicit info
Attribute 5	skimming
Attribute 6	summarizing
Attribute 7	inferencing
Attribute 8	applying background knowledge
Attribute 9	inferring major ideas or writers purpose

Procedure

The study was carried out in three stages; 1) Developing the reading comprehension test based on a cognitive assessment framework; 2) Constructing and validating a Q-Matrix of reading attributes; 3) Statistical analysis of data. The first stage of the study was to carry out an extensive study on the literature pertaining to cognitive diagnostic assessment and test development for the purpose of developing a reading comprehension test based on a cognitive diagnostic framework. From the review of literature, an initial conceptualization of test specifications based on Evidence Centered Design put forth by Mislevy (1996) was put to use in developing the 20 items for the test. After the test was developed, it was administered to 1986 students in general English courses at the University of Tehran. Data from this test administration was used for statistical analysis. The next phase was constructing a Q-matrix of L2 reading attributes. In order to construct the Q-matrix, initially a list of L2 reading attributes was specified based on the previous literature, participants think-aloud verbal reports, and content experts' judgment. The final phase included empirical validation of the Q-matrix through Fusion model analysis. Reading test data were analysed along with the Q-matrix using the Arpeggio suite software, which implements the Fusion model.

Instruments

The test used was a reading comprehension test developed for this study. The test was developed based a cognitive diagnostic framework, including three different passages along with 20 items. The three passages covered topics from natural sciences, engineering and the humanities.

Data Analyses

For the process of test development and Q-matrix construction, both qualitative and quantitative analyses were carried out. Qualitative analyses were carried out to specify reading skills assessed by the reading test. For this means, various taxonomies of reading skills and strategies in the literature were studied. Then, think-aloud verbal protocols were analyzed qualitatively to help understand the characteristics of the cognitive processes and skills used by the students and to identify primary reading skills. Six rater's ratings were also used to examine to what extent the specified skills are necessary to correctly answer the test items.

The Q-matrix was refined through Fusion Model analysis. Reading test data were analyzed together with the Q-matrix using the Arpeggio Suite software (DiBello & Stout, 2008b), which implemented the Fusion model. The first step in the Fusion model analysis is the analysis of the Markov Chain Monte Carlo (MCMC) convergence to guarantee that model parameters had a stable value (Roussos et al., 2007). Markov Chain Monte Carlo (MCMC) convergence checking is another step of the statistical analysis. Arpeggio software uses a Bayesian approach with a Markov Chain Monte Carlo (MCMC) algorithm. "The MCMC estimation provides a jointly estimated posterior distribution of both the item parameters and the examinee parameters, which may provide a better understanding of the true standard errors involved" (Patz & Junker, 1999). MCMC convergence is mainly evaluated by visually examining the time–series chain plots and density plots. With the fusion model, MCMC chains of simulated values are generated to estimate all the parameters (Li & Suen, 2013). Among the different parameters, first the convergence of examinees' probability of mastery for each attribute (p_k) was evaluated overall. Also, three parameters that indicate the item difficulty *(π_i*)*, item discrimination power *(r_{ik}*)*, and item completeness (c_i) were evaluated. In this section, examinees' L2 reading performance on the reading comprehension test was evaluated in terms of their mastery and non-mastery of L2 reading attributes.

Fit statistics were measured to evaluate the fit of the model to the data. The two types of fit statistics measured are FUSIONStats and IMStats, or item mastery statistics. The first compares the difference between the proportion of observed correct items and the proportion of estimated correct items. A low difference between the two *p*-values indicates a good fit of the data. IMStats indicate a comparison of the observed performance of masters and non-masters at the item level. Also, the reliability of the Fusion model was examined by analysing the Correct Classification Rate (CCR), which is the consistency of classification of examinees into masters versus non-masters of attributes. In the final step, the examinees' strengths and weaknesses in L2 reading ability at the attribute level were evaluated through probability of mastery for each attribute (p_k). A summary of Fusion Model analysis is provided in Table 2 as follows.

Table 2. Steps in Fusion Model Analysis (*adapted from Kim, 2011, p.103)*

Steps	Purpose	Analysis
Analysis of MCMC convergence	To ensure that the estimated parameters are stable and are ready to be evaluated	Examine plots (Chain plots/ Density plots)
Analysis of Model Fit	To evaluate the fit of the model to the data	Examine (FUSIONStats/ IMStats)
Analysis of Correct Classification Reliability (CCR)	To evaluate the rate of classification to master and non-masters of attributes	Examine CCR index
Analysis of Examinee parameters	To evaluate examinees L2 reading performance	Analyze examinee parameters Overall groups attribute mastery probability: p_k Individual attribute mastery probability: $p_k\ (X_j)$
Analysis of Item parameters	To evaluate item parameters and the quality of the reading comprehension test items	Analyze item parameters Item difficulty: π* Item discrimination: r_{ik}*

Results

Q-Matrix Development

Results from think-aloud verbal protocols and content raters judgment were analyzed to develop the list of reading attributes. This list of reading attributes was then used to develop the initial Q-matrix. As shown in Table 3, the rows of the Q-matrix indicate the 20 items from the reading comprehension test, and the columns indicate the nine reading attributes.

Table 3. Q-matrix of Attributes

Item/Attribute	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*	*A7*	*A8*	*A9*
Item1	1	0	1	0	1	0	1	0	0
Item2	1	0	0	0	1	0	1	1	0
Item3	0	0	1	0	1	0	1	0	1
Item4	1	0	0	1	1	0	0	0	1
Item5	1	0	1	0	1	1	0	0	0
Item6	0	0	1	0	1	1	0	0	1
Item7	1	0	1	0	1	0	1	0	0
Item8	1	0	1	0	1	0	0	1	0
Item9	1	0	1	0	1	0	1	0	0
Item10	1	0	1	0	1	0	0	1	0
Item11	1	1	1	0	1	0	1	0	0
Item12	1	0	1	0	1	0	1	1	0
Item13	1	1	1	0	1	0	0	0	0
Item14	1	0	1	1	1	0	1	0	1
Item15	1	0	1	0	1	0	1	0	1
Item16	1	0	1	0	1	0	0	1	0
Item17	1	0	1	1	1	0	1	1	0
Item18	1	0	1	1	1	1	1	0	1
Item19	0	0	1	0	1	0	1	0	0
Item20	1	0	0	1	1	1	1	0	1

Attributes that three or more raters agreed upon were considered as essential for the item and were included in the initial Q-matrix. According to Hartz (2002), those attributes that were measured by fewer than three items do not provide statistically meaningful information and therefore can be merged with similar attributes or deleted from the Q-matrix. The nine attributes obtained include deducing meaning from context, determining word meaning out of context, comprehending text-explicit info, comprehending text-implicit info, skimming, summarizing, inferencing, applying background knowledge, and inferring major ideas or writers purpose. Many reading attributes are involved in completing each item due to the complicated nature of reading (Alderson, 2000; Urquhart & Weir, 1998), as in this study.

Fusion Model Analysis

To statistically examine the identified attributes in the initial Q-matrix, Fusion model analysis was conducted using Arpeggio software. In order to examine the fit of the Fusion model to the data, two types of goodness of fit measures were used: (1) FUSIONStats and (2) item mastery statistics (IMStats). FUSIONStats compare the difference between the observed item p-values (the proportion of observed correct items) and the estimated item p-values (the proportion of estimated correct items). A low difference between the two *p*-values suggests a good fit of the data. The absolute difference between each observed item p-value and the estimated item p-value for each item should be below the suggested value of .05 for all items as put forth in (Roussos et al., 2007). However, in this study, this value was higher than .05 for four items out of twenty, including item 8 at .09, item 10 at .06, item 14 at .06 and item 20 at .08. Moreover, the mean absolute difference between the p-values was low at .04. Figure 1 graphically depicts the results of observed versus estimated item p-values, which suggested that the Fusion model fit well to the data.

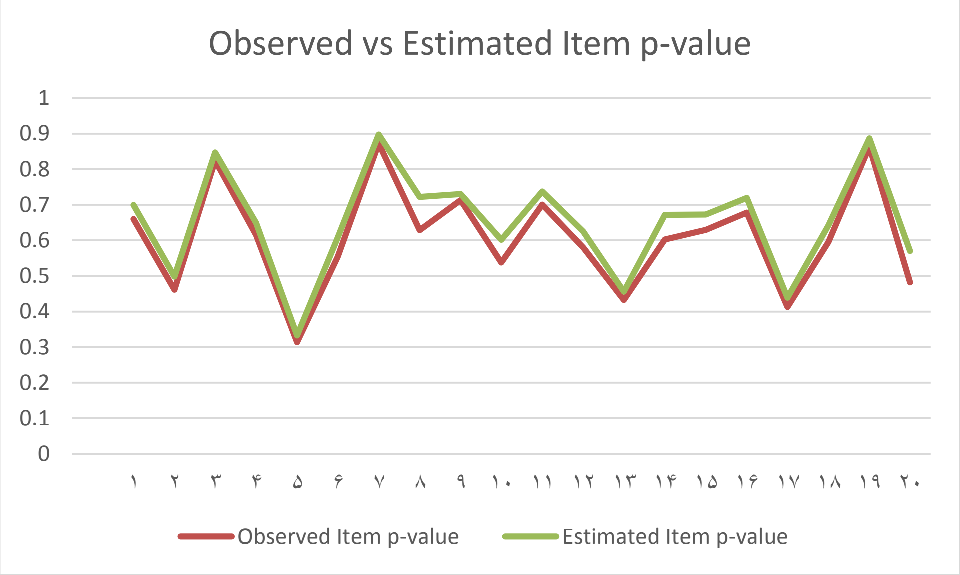

Figure 1. Observed vs. Estimated Item p-value

In addition, item mastery statistics (IMStats) were used to compare the observed performance of masters and non-masters at the item level. Three different values are considered to evaluate IMStats: *phat (m)*, which refers to the probability of correctly responding to an item given the mastery of the attributes for that item; *phat (nm)*, which refers to the probability of correctly responding to an item given non-mastery of the attributes required for that item, and *pdiff*, which indicates the average difference between *phat (m)* and *phat (nm)* across items. In this study, as shown in Table 4, the average *phat (m)* across all items was 0.774, which indicates that the average probability of getting a correct response to an item by masters of attributes was relatively high at 77.4%. On the other hand, the average *phat (nm)* was 0.322, indicating that the average probability of having a correct response to an item by non-masters was much lower at 32.2%. Thus, the *pdiff* was 0.452, indicating that the masters of attributes on an item outperformed non-masters of attributes on average by 45.2%. This high value shows a good fit between the estimated model and the observed data, indicating the strong diagnostic power of the model.

Table 4. Probability of Correctly Responding to an Item

Item	***phat(m)***	***phat(nm)***	***phat(m) – phat(nm)***
Item 1	0.841	0.398	0.443
Item 2	0.543	0.398	0.214
Item 3	0.957	0.609	0.348
Item 4	0.772	0.380	0.392
Item 5	0.358	0.197	0.161
Item 6	0.788	0.278	0.51
Item 7	0.982	0.565	0.417
Item 8	0.958	0.181	0.777
Item 9	0.960	0.312	0.648
Item 10	0.718	0.209	0.509
Item 11	0.919	0.334	0.585
Item 12	0.780	0.128	0.652
Item 13	0.552	0.169	0.383
Item 14	0.864	0.091	0.773
Item 15	0.812	0.368	0.444
Item 16	0.790	0.426	0.364
Item 17	0.513	0.301	0.212
Item 18	0.751	0.393	0.358
Item 19	0.938	0.753	0.185
Item 20	0.684	0.012	0.672
Average	0.774	0.322	*pdiff* = 0.452

In addition, the reliability of the Fusion model was examined by evaluating the Correct Classification Rate (CCR) index. The output file from Tabulator, called *classfile.csv*, reports the estimated correct classification rate (CCR) for each skill (as shown in Table 5). CCR refers to the consistency of classification of examinees into masters versus non-masters of attributes of the same test that was administered to the same group of examinees multiple times (Roussos et al., 2007). The CCR ranges between zero and one. In this data set, the CCR was high at 0.826, indicating a high reliability of the Fusion Model.

Table 5. Correct Classification Rates for Masters vs Non-Masters across Attributes

Attribute	Overall CCR (%)	CCR for M (%)	CCR for NM (%)
1	83	93	47
2	83	90	67
3	85	92	70
4	71	86	46
5	85	93	63
6	74	91	34
7	77	97	17
8	70	91	28
9	76	96	20

(M=Masters, NM=Non-masters)

Examinee parameters were analyzed to investigate test-takers' performance on the reading test in terms of their mastery probability of the L2 reading attributes in three reading proficiency groups (i.e., beginner, intermediate, and advanced) and the overall group. The L2 reading attribute mastery probability of the three levels was obtained by averaging each groups $p_k(X_j)$. The $p_k(X_j)$ value refers to the probability of mastery of an L2 reading attribute (k) on an individual examinee (*j*) level, given the item response string of the examinee *(X_j).* Each *$p_k(X_j)$* value ranges between zero and one. A value close to one indicates that the examinee has a good command of the L2 reading attribute, whereas a value close to zero suggests the opposite (Kim, 2011). Table 6 gives a brief overview of the mastery probabilities for the three levels of proficiency and the overall group.

Table 6. L2 Reading Attribute Mastery of Reading Proficiency Groups

Level	**Attributes**								
	A1	A2	A3	A4	A5	A6	A7	A8	A9
Beginner	0.23	0.053	0.16	0.112	0.154	0.065	0.088	0.06	0.098
Intermediate	0.908	0.545	0.797	0.674	0.699	0.645	0.533	0.621	0.676
Advanced	0.996	0.945	0.965	0.943	0.977	0.923	0.987	0.956	0.984
Overall group	0.769	0.676	0.662	0.627	0.723	0.699	0.754	0.665	0.737

As it is observed from Table 6, the beginners' attribute mastery probabilities were very low with mastery probabilities ranging from .053 (determining meaning out of context) to .160 (comprehending text explicit info). The intermediates' attribute mastery probabilities had a wider range with mastery probabilities ranging from .533 (inferencing) to .908 (deducing meaning from context). Naturally, the advanced learners' attribute mastery probabilities were overall very high with mastery probabilities ranging from .923 (summarizing) to .996 (deducing meaning from context). The results indicated that the advanced group performed remarkably well on the attributes with over 96.4% of advanced learners having mastered each attribute.

In order to investigate if there were statistically significant differences in the L2 reading attribute mastery probabilities among the three groups, a one-way analysis of variance (ANOVA) was conducted. Results indicated that the attribute mastery probabilities significantly differed at the $p<.05$ level for the three reading proficiency levels. Figure 2 below visually compares the performance of the three proficiency groups and the overall group. As it is observed from Figure 2, the beginners showed poor performance on the reading test with very low attribute mastery probability across the nine attributes. Their mastery probabilities showed variability among the knowledge related attributes such as deducing meaning from context with a mastery of 0.23 to determining meaning out of context with a mastery of 0.053. The fact that they had the highest probability for the first attribute is in accordance with our previous findings indicating that deducing meaning from context was the easiest with regard to difficulty.

The intermediates demonstrated average performance across attributes, as anticipated. They showed much higher variability among attributes than the other two proficiency groups, ranging from a low of 0.533 for the seventh attribute (inferencing) to a high of 0.908 for the first attribute (deducing meaning from context). Not surprisingly, the average mastery of attributes was in the range of approximately 68% for the intermediate test-takers, which is above average and acceptable. Overall, the intermediates performed much better than the beginners across attributes, while the two groups did not show similar mastery patterns, since the intermediate group showed higher variability among attributes.

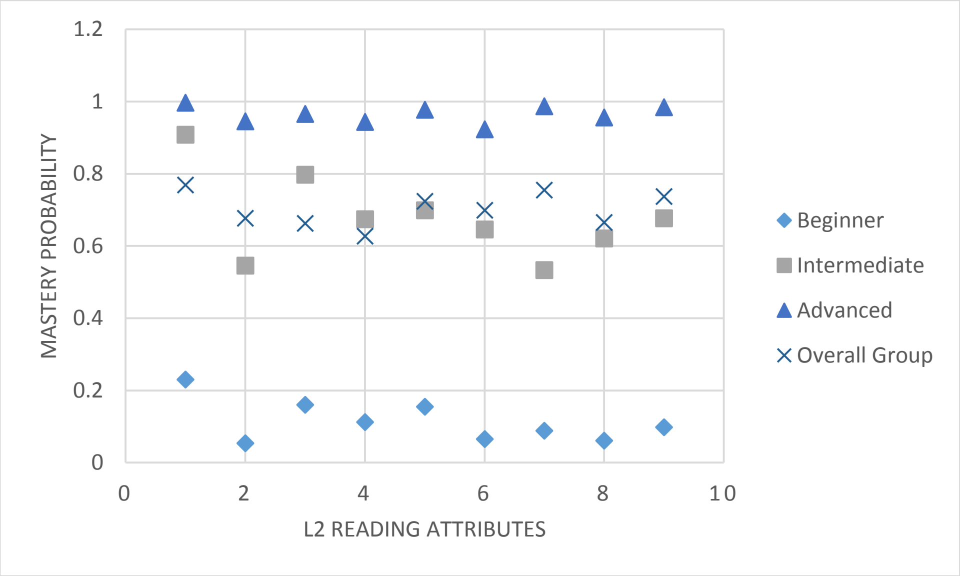

Figure 2. L2 reading attribute mastery statistics of different proficiency groups

The advanced group had very high mastery probability across all nine attributes. There was basically insignificant variability among the attribute mastery probabilities. Their probabilities were all above 0.92, indicating that they all had mastery of the L2 reading attributes with an average of 96.4 percent, which is exceptional for this test. Among the mastery probabilities, it appears that the advanced group had high mastery on the first reading attribute, deducing meaning from context, similar to the first two proficiency levels. However, the mastery probability for the seventh reading attribute, inferencing was high at 0.987, which is contrary to the intermediate group with a low of 0.533. This is clear indication of the strengths and weaknesses in diagnostic assessment, and can be very beneficial output for the teacher to consider the weakness of the intermediate group, regarding specific skills and attributes, in order to facilitate the teaching process. By observing the weak attributes from the mastery patterns, teachers can get an idea of which skills the students need to work on and hence proceed with providing new lesson plans and new instructional materials to improve those skills that the students are lagging behind on. This is the climax of cognitive assessment for the teacher, when he/she can without hesitation, put less focus on mastered attributes and allocate more class time to those skills that need greater attention.

Now looking back at the variability in mastery for attribute 7 (inferencing), the results are quite in line with previous research on L2 reading ability. According to Kim (2011), previous research indicates that advanced readers are commonly better at using inferencing strategies compared to beginners. Kim (2011) observed that advanced readers focused on the meaning of the text as a whole and used background knowledge to make inferences from the text. On the other hand, beginners depended on using single skills, such as decoding to comprehend the text better. Therefore, the statistical results from this study support findings of the previous studies especially that advanced learners had high mastery of comprehending text implicit information and inferencing, which required understanding the implied meaning of the passage.

Discussion and Conclusion

Among the four skills of English language proficiency, reading ability might well be considered an essential skill for success in the academic world. Hence, it is crucial to accurately assess learners' L2 reading ability to gauge their learning and help enhance their reading skills. The main goal of the current study was to develop a test based on a cognitive diagnostic framework in order to diagnose learners' strengths and weaknesses in L2 reading ability, with the ultimate goal of providing detailed information that can assist teachers and administrators for instructional purposes and to improve student performance.

In fact, two elements were considered in the course of this study. First was investigating the L2 reading attributes necessary for successfully completing each item on the reading test. Raters identified the various attributes, such as knowledge and strategies, by referring to a list of L2 reading attributes and students think aloud verbal reports. This list of nine reading attributes was organized into an item-by-attribute Q-matrix. Second, test-takers' performance on the reading test was examined for diagnostic purposes. The test scores were then analysed in conjunction with the Q-matrix, using the Fusion model analysis.

Findings of the study suggest that a number of items on the test can successfully discriminate between masters and non-masters and are therefore appropriate for CDA. In addition, the list of L2 reading attributes could be used as a framework for CDA research in the future. Regarding the frequency of reading attributes for each item on the test, the Q-matrix was examined to identify a recurring pattern among the L2 reading attributes. Since language knowledge and strategic competence are believed to interact with each other during the mastery of reading ability (Bachman & Palmer, 1996), it was expected that each item on the reading test would measure at least one knowledge-related attribute and one strategy. In fact, it was found that almost all items measure one knowledge-related attribute and a minimum of one reading strategy. Since strategies, by definition, manage the use of language knowledge (Bachman & Palmer 1996), they were assumed to release knowledge-related attributes.

Examinees' performances on the attributes existing in reading comprehension items were evaluated. As Roussos et al. (2007, p. 293) put forth, "A key issue for mastery/nonmastery of diagnostic models is whether the proportion of examinees estimated as masters on each skill is relatively congruent with the user's expectations." Fusion model analysis was carried out to obtain relationships among the participant's performances on the test items. The Arpeggio suite software provides a number of output files that give specific information regarding the examinees performance on each item of the test. Two of the output files that help us respond to research question 2 are the classification file (classfile.csv) and the fit report file (fitreports.csv). The fit report file is an output file that provides fit statistics. The classification file indicates the consistency of classifying the examinees in terms of their mastery or non-mastery of each attribute (i.e., the correct classification rate). First of all, the model fit the data well since mean absolute difference between the p-values was low at .04 as indicated by observed and estimated p-values. It was indicated that masters of attributes outperformed non-masters of attributes by 65.2%, which is desirable. In addition, the average *phat (m)* across all items was .774, indicating that the average probability of getting a correct response to an item by masters of attributes was relatively high at 77.4%. The average *phat* *(nm)* was .322, so the average probability of having a correct response to an item by non-masters of attributes was much lower at about 32.2%. In addition, the *pdiff was* .452, indicating that the masters of attributes on an item outperformed non-masters of attributes on average by 45.2% across all items. This high value indicated a good fit between the estimated model and the observed data, suggesting a strong diagnostic power of the model.

According to Lumley (1993), identifying implicit information (equivalent to inferencing) and synthesizing to draw a conclusion (equivalent to summarizing) were difficult compared to vocabulary (similar to identifying word meaning) and identifying explicit information (similar to finding information and skimming). This could be attributed to the fact that inferencing and summarizing are higher-level strategies involving more complex cognitive processing than the other three strategies, which require lower-level strategies as was the case in this study. As an example, summarizing requires readers to first comprehend the overall text and then extract the gist from it. Understanding the gist involves numerous components, such as knowledge of grammar, vocabulary, discourse structure, and various cognitive processes (Birch, 2002). Thus, the nature of summarizing seems quite complex. In a similar vein, the strategy of inferencing has long been believed to be a challenging one (Fletcher, 2006). In order to make inferences, readers should already have the ability to understand the literal meaning of the text, which makes inferencing more difficult to master. However, the results of this study are contrary to this belief, whereas for the overall group, the attribute of inferencing stands in second at .754 with less difficulty. On the other hand, between skimming and summarizing, the latter was more difficult with .698 as compared with skimming at .723. In fact, as put forth by Urquhart & Weir (1998), skimming involves quickly understanding the surface-level propositional meaning of the text and is considered a less challenging strategy. So while the range of difference is not very notable between these two strategies, skimming was identified as the easier L2 reading strategy in the context of the reading test.

Furthermore, comprehending text implicit information, comprehending text explicit information and applying background knowledge were considered the more difficult attributes with .627, .662 and .666, respectively. Similarly, deducing meaning from context the easiest attribute, which is in accordance with the belief that word recognition, which is similar to identifying word meaning, involves lower-level processing (Alderson, 2000). Overall, examinees have performed comparatively well on these three attributes (i.e., deducing meaning from context, inferring major ideas and skimming) due to the nature of the attribute, which requires relatively less cognitive processing.

This study also involved investigated the strengths and weaknesses of examinees in three different L2 reading proficiency groups: beginner, intermediate, and advanced. Results from the Fusion model analysis indicated that each reading proficiency group demonstrated different mastery patterns of the L2 reading attributes. Beginners had very low attribute mastery probabilities across all items; less than approximately 23% of the beginners had mastered each attribute. Intermediates performed much better than the beginners, but demonstrated a wide range of L2 reading attribute mastery probabilities, ranging from approximately 53% (inferencing) to 90% (deducing meaning from context). Contrarily, advanced learners had very high mastery probabilities of all L2 reading attributes, with mastery probabilities fluctuating above approximately 92%. In fact, an average of 96% of all advanced learners had mastered each L2 reading attribute. Comparing the attributes among the three proficiency groups and the overall group, attribute one, deducing meaning from context had the highest mastery among all groups, while the most difficult attribute varied among these different levels of proficiency. The beginners had more difficulty with determining meaning out of context, while the intermediate group scored less on the inferencing attribute. Advanced learners had more difficulty with summarizing, while the overall group had difficulty with comprehending text implicit information.

Thus, based on the three reading proficiency group's attribute mastery patterns, it was possible to infer their strengths and weaknesses in L2 reading. Beginners and intermediates' mastery of attributes were disproportionate. That is, they had high mastery over certain attributes, but lower mastery over others, clearly demonstrating their strengths and weaknesses. Among the knowledge-related attributes, beginners and intermediates had the lowest mastery in pragmatic meaning, indicating that it was their weakness. On the other hand, the two groups had the highest mastery in lexical meaning, suggesting that this was their greatest strength. Among the strategies, they had lower mastery of summarizing and inferencing, but noticeably higher mastery of deducing word meaning and skimming. Thus, summarizing and inferencing were their weaknesses, while deducing word meaning and skimming were their strengths. Advanced learners showed high mastery probabilities over all knowledge-related attributes and strategies; that is, they excelled in all attributes and did not appear to have specific weaknesses in reading.

The detailed score reports of this study provide beneficial in facilitating learning on the part of the student and in teacher preparation and curriculum development on the part of the teacher. With detailed reports of test results, teachers can become aware of students' problematic areas and focus on them in lesson planning and providing learning material. Since the reading test developed here was based on a cognitive framework followed by Fusion model analysis, the problematic items were identified and could be further modified and replaced. Hence, an item bank can be developed for cognitive diagnostic development items in the current test and those in future studies.

Limitations and Suggestions for Future Research

The current study provided a number of implications for teachers and practitioners, both pedagogically and theoretically. First and foremost is the pedagogical implication for assessment purposes in an attempt at constructing a cognitive diagnostic test to gauge L2 reading proficiency. The need for developing tests based on cognitive diagnostic frameworks is becoming more and more necessary in the realm of assessment and language testing. The type of diagnostic feedback that can be provided to teachers includes attribute mastery probability for overall groups, different proficiency levels, and individual students. Teachers can refer to this information to refine and upgrade the L2 reading material essential to meet the needs of each proficiency level.

As the process of all research faces some limitations, the present study might also suffer from some. While determining the attributes of a Q-matrix necessitates a deep understanding of the nature of cognitive skills and items, the complexity of the reading skill does not allow a full understanding of its cognitive processes (Lee & Sawaki, 2009). In addition, there is a lack of consensus on the skill components of reading comprehension (Alderson, 2000). Hence, although a great number of attributes may be identified, not all attributes can be kept in the Q-Matrix for Fusion Model analysis. As a result, the purpose is not to identify all the attributes that could be involved in responding to the reading comprehension items, but to examine the major attributes required to successfully complete each item. Therefore, the reading attributes are not exhaustive, but are specifically related to the reading comprehension test. Meanwhile, those attributes that were identified for this study were not all appropriate for parts of the Q-matrix, specifically for items 15-20. Thus, exploration of additional attributes to be used for Q-matrix construction is recommended for future studies.

In addition, the reading comprehension test, which was devised based on a cognitive diagnostic framework, is only at the beginning stages of experimentation. Further research should be carried out to devise more cognitively based tests, not only for the reading skill, but also for writing and listening assessments. Devising cognitively-based assessments requires numerous pilot studies and carrying out further research can improve the accuracy of items developed.

One of the limitations of the study was the format of the reading comprehension test, which was in the form of strictly multiple choice items. In future studies, other item types such as fill-in-the-blank, essay type and open-ended questions should be considered. Also, a suggestion for future research is to focus on item distractors to enhance the diagnostic potential of the test. The proportion of diagnostic information that is obtained from such a test greatly depends on the test design and item construction. Specifically, in a retrofitting approach, not all test items can diagnostically differentiate examinees based on their underlying skill competencies. However, test/item diagnostic discrimination can be enhanced by observing incorrect response patterns of examinees. This necessitates developing diagnostically sensitive distractors in MC items (Gu, 2011).

Another limitation of the study is that the Q-matrix, which is essential in the Fusion model analysis, could be developed by a greater number of content experts, including training sessions prior to rating. In the current study, six context experts rated the L2 reading attributes; which is sufficient for the purpose of the study, but having a greater number of experts could possibly make the Q-matrix more productive. Since the quality of the Q-matrix determines the quality of the Fusion model analysis, it is important to take extreme precaution in developing it (Jang, 2005).

While this study was an attempt at diagnostic assessment of L2 reading attributes, it also demonstrates a need for continued research in the area of cognitive diagnostic assessment in the Iranian context, particularly with regard to constructing diagnostic tests for different skills including writing, speaking and listening. Therefore, now is the time to focus our attention on designing and developing educational assessments that are based on a CDM framework. Carrying through such an endeavor necessitates the cooperation of various experts from different fields (i.e., subject matter, learning sciences, measurement, and pedagogy). By succeeding in such an effort, educational assessments will become more instructionally-oriented and more relevant to the needs of present day classrooms.

References

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

Alderson, J. C. (2010). “Cognitive diagnosis and Q-Matrices in language assessment”: A Commentary, 96-103.

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

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 of Research in Mathematics Education, 24,* 442-459.

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.

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

De la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163-183.

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

DiBello, L., & Stout, W. (2008). Arpeggio documentation and analyst manual. *Chicago: Applied informative assessment research enterprises (AIARE)—LLC.*

DiBello, L., & Stout, W. (2008). Arpeggio suite, version 3.1. 001 [Computer program]. *Chicago: Applied informative assessment research enterprises (AIARE)—LLC.*

Embretson, S.E., Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38(4),* 343-368.

Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading*, *10*(3), 323-330.

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.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 7(2),* 145-220.

Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 63(2-B), 864.

Hartman, H. J. (2001). Developing students' metacognitive knowledge and skills. In H. J. Hartman (Ed.). *Metacognition in learning and instruction: Theory, Research and Practice* (pp. 33-68). Dordrecht, the Netherlands: Kluwer.

Huang, T.W., & Wu, P.C. (2013). Classroom-based cognitive diagnostic model for a teacher-made fraction-decimal Test. *Educational Technology & Society, 16* (3), 347–361.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL.* Available from ProQuest Dissertations and Theses database. (AAT 3182288)

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, *6*(3), 210-238.

Kim, A. Y. (2011). *Examining second language reading components in relation to reading test performance for diagnostic purposes: A Fusion model approach* (Doctoral dissertation, Teachers College, Columbia University).

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*. [0265532214558457].

Lee, Y. W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*(3), 239-263.

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.

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

Li, H. (2011). Evaluating language group differences in the subskills of reading using a cognitive diagnostic modeling and differential skill functioning approach [Doctoral dissertation, The Pennsylvania State University].

Li, H., & Suen, H. K. (2013). Constructing and Validating a Q-Matrix for Cognitive Diagnostic Analyses of a Reading Test. *Educational Assessment*, *18*(1), 1-25.

Lumley, T. (1993). The notion of sub-skills in reading comprehension test: An EAP example. *Language Testing, 10(3),* 211-234.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*(4), 379-416.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439–483.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing. Special Issue: Interpretations, Intended Uses, and Designs in Task-based Language, 19*(4), 477–496.

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. *Educational measurement*, *4*, 257-305.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics,* 24, 342–366.

Pellegrino, J. C., & Chudowsky, N. N. & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* (National research council's committee on the foundation of assessment) National Academy Press; Washington D.C.

Roussos, L. A., DiBello, L. V., Stout, W. F., Hartz, S. M., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnostic system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). New York, NY: Cambridge University Press.

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., & Henson, R. A. (2012). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

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.

Sheehan, K., & Mislevy, R. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27,* 255-272.

Snow, R. E., & Lohman, D. F. (1989). *Implications of cognitive psychology for educational measurement*. American Council on Education.

Stiggins, R., Arter, J., & Chappuis, S. (2004). *Classroom assessment for student learning: Doing it right—using it well.* Dover, NH: Assessment Training Institute.

Svetina, D., Gorin, J.S. & Tatsuoka, K.K. (2011) Defining and comparing the reading comprehension construct: A cognitive-psychometric modelling approach, *International Journal of Testing*, 11:1, 1-23.

Tatsuoka, K. K. (1995). Architecture of knowledge structure and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-361). Hillsdale, NJ: Lawrence Erlbaum Associates.

Templin, J. L. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis.* [Available from ProQuest Dissertations and Theses database. (AAT 3160960)]

Urquhart, S., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice.* New York: Longman.