Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 7, No. 2, October 2017
ISSN 2476-5880

Cognitive Diagnostic Assessment: Issues and Considerations

Zahra Javidanmehr[1], Mohammad Reza Anani Sarab[2],

Received: 11 March 2017 Accepted: 25 August 2017

Abstract

Cognitive Diagnostic Assessment (CDA) is a type of educational assessment that is designed to measure specific knowledge structures and processing skills in students so as to provide information about their cognitive strengths and weaknesses (Leighton & Gierl, 2007). CDA has been instrumental in turning the attention of practitioners to more diagnostic, descriptive, and fine-grained levels of feedback. Different statistical, psychometric models, called Cognitive Diagnostic Models (CDMs), have been proposed to extract this kind of information from both diagnostically and non-diagnostically designed tests. These models provide two sets of information to the test users: information on mastery/non-mastery patterns of sub-skills for test-takers and information on the diagnostic power of test items. Due to its novelty and relative complexity of its procedures, cognitive diagnostic assessment is still far from achieving its proper place in educational assessment. This paper aims at providing an easy-to-grasp account of CDA's theoretical foundation and its procedures of test analysis. The present paper first focuses on what and why of CDA in education and second language acquisition. In this part, theoretical underpinnings of CDA, psychometric models of practicing the analyses, model selection, and studies in SLA are presented. The second section presents how these foundations are put into practice in a stepwise manner. Four main steps are delineated in conducting a CDA analysis. The procedural steps are then exemplified using real data for analysis. The paper concludes with an account of the limitations and untapped areas in CDA.

Keywords: *cognitive diagnostic assessment, educational measurement, high stakes testing, Q-matrix, reading comprehension*

1. Introduction

With the burgeoning studies on the impact of testing on society, education system, and individuals (Messick, 1989; Shohamy, 1992), the educational testing community has increasingly felt the limitations of the output of the current testing practices, which rank order test-takers based on their test performance. Score reporting of current testing was rightly

1 ShahidBeheshti University, Tehran, Iran. *(corresponding author)* Email: javidanmehr@yahoo.com

2 ShahidBeheshti University, Tehran, Iran. Email: reza_ananisarab@yahoo.co.uk

challenged in the mid-1980s. Splolsky (1990), for instance, recommended using profiles in which multiple sub-skills are reported in more than one way. In the same vein, Shohamy (1992) called for detailed and diagnostic method of feedback reporting and some other arguments were also made in favor of more descriptive test information in order to improve instructional design and guide students' learning. Consistent with these arguments, Cognitive Diagnostic Assessment (CDA) was introduced as a new method in educational measurement that can provide fine-grained diagnostic information about test-takers' degree of mastery of some domain sub-skills (Lee & Sawaki, 2009). Sub-skills are defined as domain-specific knowledge and skills that are required to indicate mastery in a specific cognitive domain (Leighton & Gierl, 2007). Taking reading skill as a cognitive domain, one needs to have knowledge of vocabulary, grammar, making inferences and so on in order to comprehend a text completely. These are considered the sub-skills of the reading domain, which are called attributes as well. Throughout this paper, these two terms are used interchangeably. The most conspicuous characteristic of this approach is that it is the point where cognitive psychology and psychometric modeling meet within a single framework, therefore it aims to assess the test-takers" knowledge and underlying cognitive processing sub-skills (DiBello, Roussos, & Stout, 2006; Leighton & Gierl, 2007).

The demand for CDA can partly be explained by the fact that the traditional forms of assessment have failed to satisfy the expectation of diagnostic feedback. Both Classical Test Theory (CTT) and Item Response theory (IRT) locate test-takers on a trait scale by providing only an overall score of their proficiency level in the target domain (Choi, Rupp, & Pan, 2012). According to Rupp, Templin, & Henson (2010), the score that is reported in this way is inadequately beneficial in supporting formative interpretations for qualitative diagnostic purposes. In comparison, a test designed based on CDA principles is capable of specifying the test taker's latent proficiency level score along with an indication of its underlying knowledge structures. This specification allows for possible intervention to address individual and group needs and improve instruction for students' effective learning and progress (Lee, 2009).

Due to the diagnostic nature of CDA, it not only can reveal more precise and detailed information to the users of a test, but also can disclose the test-takers' weaknesses and strengths on the pre-specified sub-skills of the target domain (Leighton & Gierl, 2007; Rupp *et al.*, 2010). CDA can be considered as a step forward regarding the feedback that educational measurement has always strived to provide. This goal cannot be achieved unless CDA is operationalized and practiced both in classroom assessment and large scale testing. The major hurdle in the way of achieving this goal is the fact that CDA is still in its infancy and although its foundational theories are well-established, its operationalization is still on the way. The complexity of the statistical methods and the interpretation of their outputs have limited the expansion of CDA into mainstream assessment.

Therefore, the main purpose of this article is to provide an easy-to-grasp scheme for the practice of CDA for those who are new to the area or are interested in this approach to assessment. Attempts have been made to sketch CDA without reference to mathematical jargon as to make the concepts and procedures more accessible to practitioners. This paper includes two major parts. In the first part, theoretical underpinnings of CDA are presented. This section includes a short historical background of diagnostic assessment, limitations of traditional

approaches, statistical models of analysis in CDA, and CDA in second language acquisition (SLA). In the second part, four main steps in conducting experimental cognitive diagnostic analysis are described in extensive detail. The steps are then exemplified using real data. Some limitations of CDA practice and untapped areas come next.

2. Theoretical Underpinnings of CDA

2.1. Historical background

CDA can be traced back to the writings of scholars such as Messick (1989) on test validity and Snow and Lohman (1989) on cognitive psychology. Not directly mentioning CDA, Messick (1989) highlighted the significance of inferring information about test-takers' mental processes from the scores they get on a test. Nichols (1994) and Mislevy, Nichols, Chipman, and Brennan (1995) thereafter coined the term *cognitively diagnostic assessment* to refer to implementing cognitive diagnosis in the context of education (Leighton & Gierl, 2007). CDA, identified at the inter-section of *cognitive psychology*, which studies mental representations of human's observable behaviors, and *psychometrics*, which is devoted to measuring skills, knowledge, abilities, and attitudes, has attracted the attention of researchers and educational measurement students since the mid-1980s. In the last two decades, contributors from both research communities of cognitive psychology and psychometric theory have started to lay the ground for the new approach to assessment with the hope of compensating for the failure of item response theory (IRT) and classical test theory (CTT) to provide diagnostic information to supplement test results.

The mission of these traditional psychometric measurement models has been rank-ordering test-takers on an underlying latent trait (construct) and locating them in a group of examinees. In IRT, for example, the relationship between an examinee's responses to test items and a latent variable is specified by a mathematical function, and the test result for the examinee produces a single score as a measure of an underlying latent variable (Hambleton, Swaminathan, & Rogers, 1991). Although these test results are valuable for ranking and comparing examinees, Snow and Lohman (1989) conclude that they lack a substantive psychological theory in order to explain item responses through explicit demonstration of the psychological processes that underlie the test constructs. Furthermore, these models reflect the investigators' expectations of how students will come to an answer in test situations. The actual thinking procedures on the part of the students are not empirically examined (Nichols, 1994). As Jang (2005) mentions, in these assessment approaches, understanding and interpreting the meaning of an examinee's score per se is not an easy job. No information on test-takers' strengths and weaknesses is provided in the test summary results. In order to provide more detailed information on instruction and student learning, CDA seems more appealing (Lee, de la Torre, & Park, 2012).

2.2. Cognitive diagnostic assessment

A test, informed by CDA, can specify the potential knowledge structures underlying the overall test score. This specification can function as feedback that can be used in addressing individual and group needs through remedial instruction and making improvements in instruction with the

aim of enhancing learning and advancement (Lee, 2009). The definition provided by Kubinger (2006) as translated by Rupp *et al*. (2010) offers a conceptual elucidation of CDA potential referred to above:

> Diagnostic assessment is a systematic process that seeks to obtain specific information about psychological characteristics of a person by using a variety of methods. Its objective is to justify, control, and optimize decisions and their resulting actions. This process includes (a) the specification of the diagnostic question, (b) the selection of the diagnostic methods, (c) the application and evaluation of the data from the diagnostic methods, (d) the interpretation of the data and the development of a diagnostic report, (e) the design of an intervention, (f) the evaluation of the effectiveness of the intervention (p.11).

The main focus of this definition is on the systematic procedure that CDA practitioners should follow from defining the purpose of assessment to categorizing participants according to their underlying abilities.

As CDA is fundamentally diagnostic, statistical models, called cognitive diagnostic models (CDMs), are utilized in order to provide discrete attribute profiles for test-takers, which are a series of attributes or sub-skills and their related probabilities(Rupp, *et al*., 2010, p. 83). The purpose of CDMs is to classify examinees as masters or non-masters of the predetermined sub-skills/attributes based on their observed response patterns (von Davier, 2005). Different definitions and classifications are provided for CDMs, nowhere, though one can find a more detailed definition than the one provided by Rupp and Templin (2008). To them CDMs are probabilistic, confirmatory multidimensional latent-variable models. CDMs allow for a simple or complex loading structure. They include observable categorical response variables and also unobservable (i.e., latent) categorical predictor variables. The latter are combined in compensatory and non-compensatory ways to generate latent classes.

This definition highlights the distinctive features and classification of CDMs. On the grounds that CDMs rely on multiple latent variables to classify the test-takers, they are considered multidimensional models and hence are similar to multidimensional IRT models and also multidimensional factor analysis models. The departure line between CDMs and these models is the nature of the latent variables in each model. In traditional IRT and FA models, the latent variables are continuous, however CDMs utilize categorical latent variables. To put it another way, continuous latent variables are capable of only rank-ordering the test-takers and also they do not lead to statistical classifications. In addition, FA /confirmatory FA use continuous data and IRT/multidimensional IRT use discrete data. CDMs, though, tends to be defined for categorical data. Loading structure is another distinction of CDMs and MIRT and CFA. The latter models use a simple leading structure, that is to say each item loads on one latent dimension. CDMs, on the other hand use more complex loading structure, in which each item loads on multiple dimensions (Rupp et al., 2010).

2.3. Cognitive diagnostic models (CDMs)

ISSN 2476-5880

Chronologically speaking, CDMs have been developed and progressed since the introduction of diagnostic assessment. Tatsuoka (1983) developed the rule-space methodology that was applied to mathematics assessment. It is a widely-known model in educational research which was firstly applied to a set of subtraction and addition data and afterwards was utilized to examine other content areas, including second language acquisition. Embretson's (1985) Multicomponent latent trait model, the tree-based regression method (Sheehan, 1997), the hybrid model of latent class analysis (Yamamoto & Gitomer, 1993), the cognitive design system (Embretson, 1998), thedeterministic inputs, noisy and gate (DINA) model and the noisy inputs, deterministic and gate (NIDA) model (Junker & Sijtsma, 2001), the Fusion Models (Hartz, 2002; Roussos *et al*., 2007), the attribute hierarchy method (Leighton, Gierl, & Hunka, 2004), the general diagnostic model (GDM, von Davier, 2005) were introduced subsequently. Table 1 depicts some well-known and most frequently applied cognitive diagnostic models.

Table 1

Some Cognitive Diagnostic Models and their Statistical Packages

Abbreviation	Name of the model	The developer(s)	The software package
RSM	Rule space methodology	Tatsuoka (1983)	BUGLIB/ BUGSHELL
GDM	General diagnostic model	von Davier (2006)	MDLTM
RUM/RRUM	Fusion model	Hartz, 2002; Roussos *et al*., 2007	Arpeggio
DINA	Deterministic inputs, noisy and gate model	Junker & Sijtsma, 2001	Mplus and R package
NIDA	Noisy inputs, deterministic and gate model	Junker & Sijtsma, 2001	Mplus and R package
AHM	Attribute hierarchy method	Leighton, Gierl, & Hunka, 2004	R package
G-DINA	Generalized DINA model	de la Torre, 2011	R package and Ox code

Based on the features that CDMs share, some classifications are suggested. In one classification, CDMs fall into two categories: non-compensatory and compensatory (DiBello *et al.*, 2007; Roussos *et al.*, 2007). In non-compensatory models success in an attribute does not

compensate for the deficiency in another attribute in order to correctly respond to an item. DINA model, NC-RUM/RUM models, and NIDA model fall into this category. In compensatory models on the other hand, a high level of competence in one attribute can compensate for a deficiency or low level of competence in another attribute by means of the interaction of attributes required by that task. DINO model, NIDO model, and GDM are examples of this category.

Conjunctive and disjunctive models are another grouping of CDMs. This classification prescribes how attributes are combined to produce a latent response (Rupp, *et al*., 2010). Conjunctive models assume that all attributes leading to a positive response should be taken correctly (Rupp & Templin, 2008). The implication is that a missing attribute cannot be compensated for by the mastery of otherattributes. Conjunctive models are mostly used for mathematical tests which require all attributes to perform successfully on an item (Tatsuoka, 1990). In disjunctive models though successful performance on an item only requires that a subset (in some cases only one) of the probable strategies is effectively applied (DiBello, *et al.*, 2007). Disjunctive models are appropriate when multiple strategies exist to solve the item. The DINA model, for example, is viewed as a non-compensatory model that uses a conjunctive condensation function. However, the DINO (deterministic inputs, noisy or gate) model is classified as a non-compensatory, disjunctive model.

The type of the data, that is to say the test-takers' ' scored items, is another critical factor in selecting a specific CDM. Data to be analyzed can be either dichotomous or polytomous. Most educational achievement/proficiency assessments are dichotomously scored, that is 1 for a correct answer and 0 for an incorrect one. The most common example of a dichotomous item is a multiple-choice test item, which typically has 4 to 5 options, but only two possible scores (0 or 1) can be assigned to a response to such an item. True/False or Yes/No items are other binary examples. For the very obvious reason, almost all CDMs are developed to handle these types of scored tests. Polytomous responses (either nominal or ordinal) have more than two possible scores. The most common examples are Likert-type items (rated on a scale of 1 to 5) and partial credit items (scores on an essay item ranging from 0 to 5 points). Some of CDMs have been constructed to handle both dichotomous and polytomous data. The Generalized Diagnostic Model (GDM; von Davier, 2005), the Rule Space model (RSM; Tatsuoka, 1985, Tatsuoka & Tatsuoka, 1989), and the reduced non-compensatory reparameterized unified model (NC-RUM; DiBello *et al*., 1995; Hartz, 2002) are among these models.

How to choose among a list of CDMs for a particular study is not always an easy decision. The theories on the target domain might make it obvious which models would be of help. For instance, in the application of CDMs in mathematics, where the solutions involves some specific steps in a row, it is argued that non-compensatory models are the most appropriate ones (Roussos *et al*., 2007). Regarding some other domains that are compensatory in nature, such as reading comprehension, some scholars believe that compensatory models would work better (Li & Lei, 2015). Another consideration in model selection is model complexity. Compared to the use of a complex saturated model, the use of simpler constrained models may offer much more meaningful interpretations (Rojas, de la Torre, & Olea, 2012). However, Li and Lei (2015) suggest that whenever the relationships among cognitive sub-skills are not entirely known, it is

more rational to use a saturated CDM, which is flexible to report different kinds of relationships among the related sub-skills. The log-linear CDM (Henson, Templin, & Willse, 2009), the general diagnostic model (GDM; von Davier, 2005), and the G-DINA (generalized deterministic inputs, noisy and gate)model (de la Torre, 2011) all offer a general framework that include some other constrained CDMs.

To provide assessment-related diagnostic information through CDA, one can assume two different approaches to study design: diagnostically and non-diagnostically-constructed designs. These two types of designs are presented below:

2.4. Approaches to research design

2.4.1. Diagnostically-constructed designs

The most desirable form of cognitive diagnostic assessment is the one that is diagnostically designed, constructed, and scored from the very first step. In this approach cognitive sub-skills are explicitly defined to be targeted in the test construction phase. These predetermined attributes should be in line with the instructional goals. When the sub-skills are set, the data are to be analyzed with an appropriate CDM. The scores, afterwards, are to be reported in a fine-grained diagnostic system. Although the fine grained cognitive diagnostic assessment was intended to inform instructional settings in this way, the diagnostically constructed designs have hardly been discussed in the literature. A few tests, though, have been designed in order to fulfill the needs of diagnostic analysis (e.g. DIALANG by Alderson 2005, Alderson & Huhta 2005; DELNA (www.delna.auckland.ac.nz/uoa); DELTA by Urmston, Raquel, & Tsang, 2013). The problems which prevent practicing this design will be discussed later.

2.4.2. Non-diagnostically-constructed designs

The literature is replete with reports of this design which has assumed a reverse engineering approach. It is mainly retrofitted to a set of existing proficiency/achievement tests with the intention of extracting and reporting cognitive sub-skills assumed to be measured through the test. Contrary to the previous design, sub-skills are not recognized and written down by test developers, but extracted post hoc by means of a set of tools; such as test-takers' think-aloud protocols and experts' judgements. In the next step, with the help of a selected CDM, fine-grained information on mastery classification of the test-takers is inferred. Using CDA methods in the field, researchers have reported the cognitive diagnostic results from non-diagnostically designed tests (Jang, 2005, 2009; Li, 2011). One may question the value of this type of analysis. It is, though, believed that such retrofitting efforts could serve as a critical step in advancing diagnostic second language assessment research. Before delving into an expensive, time-consuming process of designing a new diagnostic test, it is worth investigating the extent to which useful diagnostic information could be extracted from existing assessments (Lee & Sawaki, 2009).

Bearing in mind its relative new status, CDA and its corresponding statistical models have not yet found their way into educational settings. The research studies, meager in number though, have targeted both the theoretical foundations of diagnostic measurement and also the real data set analysis in order to address the practical aspects. Delving into the theoretical dimension of the CDA's literature, one can find a number of all-inclusive reviews scrutinizing CDA, CDMs, the applications, and the challenges ahead. von Davier (2009) and Rupp and Templin (2008) are the most remarkable and noteworthy among them. Some others have taken a close look at the Q-matrix, its construction methods, validation process, and also uncertainties in its application (*e.g.*, Alderson, 2010; Chiu, 2013; DeCarlo, 2012; Li & Suen, 2013; Liu, Xu, & Ying, 2012; Sawaki, Kim, & Gentile, 2009). Taking the experimental dimension into consideration, Jang's (2005, 2008, 2009) analysis of the diagnostic capacity of the reading comprehension section of LanguEdge™ test items and Li's (2011) research about MELAB (Michigan English Language Assessment Battery), both using the Fusion Model (or non-compansatory reparametrized unified model) are the most-cited ones. Other studies include Kasai (1997), Lee & Sawaki (2009), von Davier (2005), Kim (2011), Aryadoust (2011), Ravand, Barati, and Widhiarso (2012) to name a few in educational measurement.

As a burgeoning area of research in educational measurement, CDA has been examined in relation to other lines of inquiries as well. Computerized Adaptive Testing (CAT) is one of these combinations, which takes advantage of the feedbacks provided by CDA (for instance, Cheng, 2010; Liu, Ying, & Zhang, 2013; McGlohen & Chang, 2008). Differential Item Functioning (DIF) is another domain which has been recently furnished by the advantages of CDA (Li, 2011; Li & Suen, 2013). Although the studies done so far have taken the educational assessment a long step forward, CDA is not practiced to the fullest yet. Bringing the theories more into light and also highlighting the procedural steps in an easy and straightforward fashion to be grasped even by unprofessional stakeholders in psychometric models might make CDA a promised area.

In addition to the studies which have dealt with different aspects of CDA's theory and practice, there are very few studies which have tried to make the application of CDA to existing tests more accessible. Some empirical studies clarify the steps to be taken in a CDA research. Jang's (2005) is considered a study guide in this regard. She has utilized the Fusion model as the CDM and scrutinized a reading comprehension test's cognitive diagnostic information. Li (2011) and Ravand and Robittzch (2015) are other studies in this line, which applying CDMs, have focused on the procedures of CDA experimental analyses. Yet, some other studies have focused most on the procedures of CDA. Lee and Sawaki (2009), in a comprehensive descriptive review, explained the application of CDA to language assessment. The major steps in data analysis and the final diagnostic score reporting have been presented in their paper. In one section of their paper, DiBello *et al*. (2006) also describe CDA from design to scoring for dichotomous data. These articles are very influential, not ample in number though, in leading non-professional CDA researchers. Following the same order, we present the steps of a CDA analysis in the following section.

3. Practical Aspects of CDA

Before presenting the procedural steps of analysis, we should acknowledge that the application of CDA to assessment measures is not a linear process. Each individual step needs to be taken in combination with the other steps (DiBello *et al*., 2007). CDA application, irrespective of the selected approach includes four main steps:

1. Extracting/defining the target attributes

2. Constructing and validating the Q-Matrix

3. Analyzing data with a CDM

4. Report mastery classification pattern

Each step is looked upon briefly as follows:

3.1. Extracting/defining the target attributes

Based on the approach chosen, the target sub-skills are extracted or defined. The term sub-skill, which is named attribute in some studies, refers to unobservable or latent characteristics of students (Choi *et al*., 2011). Sub-skills in CDA are defined as cognitive processes, strategies, and skills that underly the test items (Lee & Sawaki, 2009; Rupp *et al*., 2010). For the first design, that is to say the diagnostically-constructed one, attributes are defined precisely by test developers based on the instructional goals of the course or the assessment. These attributes, as the baselines of Q-Matrix construction and CDA, need to be defined precisely and thoroughly.

One important issue in attribute specifications is the number of attributes for a specific test. Attributes in CDA should be defined in a detailed, fine-grain size. The grain-size of an attribute is the level of specificity with which a researcher intends to dissect a cognitive response process and describe its constituent components (Rupp *et al.,* 2101). Coarse-grained descriptions of attributes and cognitive processes are often used in tables of specifications or blueprints for educational assessments. Fine-grained attribute descriptions, on the other hand, are used in standards-based assessments with the purpose of providing descriptive feedback for instruction and assessment (Leighton & Gierl, 2007). The proper grain-size depends on the objective of the diagnostic assessment and the level of specificity with which one would like to make statements about respondents (Rupp *et al*., 2010). There is no solid regulation for the number of attribute labels for a specific test. Although in a mathematical sense, CDMs can measure an unlimited number of attributes; in a practical sense an upper limit of 10 attributes makes good sense, due to the number of possible combinations of items possible (DiBello *et al.*, 2007). It is suggested that every attribute should be assessed by at least three items. It would make the results much more interpretable.

For the second design, that is to say the non-diagnostically-constructed design, various sources in combination might inform attribute extraction. The sources as mentioned in Rupp *et al.* (2010) are as follows:

3.1.1. Verbal reports or test-takers think-aloud protocols

In this source of information, test-takers either concurrently (i.e., while they are responding to the items) or retroactively (i.e., after they have responded to the items) reveal the underlying knowledge they need in answering each item.

3.1.2. Expert panel's judgment

Another useful source of information in specifying underlying attributes comes from expert panel discussion. Truly knowledgeable experts are asked to describe the processes involved based on their research and experience in the target domain's assessment.

Not mentioned in Rupp *et al*. (2010), *test specifications (blueprints)* and *concurrent literature*, that is content domain theories on the target domain are considered and utilized in attribute specifications (Buck & Tatsuoka, 1998; Leighton & Gierl, 2007). These methods can be used separately or in combination with each other. To add to the validity and reliability of the tool (Q-matrix), scholars suggest that more than one method be used in each study.

3.2. Constructing a Q-matrix

When the attributes are recognized, the next step is constructing a tentative Q-Matrix. The Q-matrix along with examinees' scored item responses would make the input for CDM data analysis. Q-matrix, described as "a mapping structure that indicates the sub-skills required for successfully answering each individual item" (Li & Suen, 2013, p. 5). It is considered as 'the quintessential component" of all CDMs (Rupp *et al*., 2010. p. 49). In other words, the particularization of which attributes are measured by each item is done numerically in a table called the Q-matrix (Tatsuoka, 1983, 1990). A Q-matrix traditionally maps the items in the rows and the attributes or sub-skills underlying the answer to an item in the columns. 1s and 0s are the entries of the table, which indicate if an attribute is measured by an item or not. A correct answer to an item may depend on one or more sub-skills. From a statistical perspective, though, the Q-matrix is the loading matrix that illustrates which item is associated with which latent variable. Table 2 illustrates a sample of a Q-matrix for 10 items and 6 attributes (taken from Rupp *et al.*, 2010).

Table 2

The Q-Matrix for 10 Items and Attributes

ISSN 2476-5880

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
Item 1	0	0	0	1	0	0
Item 2	0	0	0	1	1	0
Item 3	1	0	0	1	0	1
Item 4	1	1	1	1	0	0
Item 5	1	0	0	0	0	1
Item 6	0	1	0	1	1	0
Item 7	1	1	1	0	0	0
Item 8	0	1	0	1	0	1
Item 9	0	1	0	1	1	1
Item 10	0	0	0	0	0	1

Note. This table shows which attributes are measured by a particular item. There are 10 items and 6 attributes on this assessment. An entry of 1 indicates that the attribute is measured by an item, whereas an entry of 0 indicates that it is not measured.

3.3. Analyzing the data and the outputs

The initial Q-matrix and the test-takers' scores are to be inserted into the statistical model data analysis tool which, based on what was previously discussed about the CDM selection, would be the most appropriate one for the study (Figure 1).

3.3.1. CDM outputs

Two main elements of any CDM statistical model: the tentative Q-matrix made based on the underlying attributes and the scored items. Different models, based on their complexities, provide various diagnostic information on both test items' diagnostic capacity and test-takers' mastery/competency level on a number of cognitive attributes. Regarding the mastery classifications, some statistical models divide the test-takers into two groups of masters and non-masters, some others, on the other hand, report three groups of masters, non-masters, and indeterminate (e.g., the Fusion Model; Hartz *et al.*, 2002). Generally speaking, every CDM is expected to offer all or some of the following types of information:

3.3.1.1. Model fit

As it was mentioned earlier, choosing the best model has always been a challenge for CDA practitioners. Selecting an inaccurate model and/or constructing a mismatched Q-matrix heavily impacts the classification accuracy of attribute mastery (Li & Lei, 2015; Rupp & Templin, 2008). In this regard, model-fit indices are supposed to help in selecting the most appropriate model for a set of data and Q-matrix.

Model-fit indices in CDA are classified into relative fit (comparing the model with other existing and rival models) and absolute fit (checking fit of the model to the data). *Relative fit* index is used for the sake of model comparisons. It is critical to estimate a relative fit index before checking the absolute one as it eliminates certain candidate models (Rupp *et al.*, 2010). Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarzer, 1976) are two main criteria for this model fit. Models that yield the smallest value for each of these relative fit indices are preferred (Li & Lei, 2015; Rupp *et al.*, 2010). Although it is

very important to touch upon relative fit index, scholars warn against the use of relative fit index as the sole measure of model-data fit, because they do not reveal whether or not a model actually fits the data in an absolute sense (Rupp *et al.*, 2010).

Absolute fit indices, on the other hand, aim to measure the magnitude of overall discrepancy between observed and model-predicted values and are typically functions of residuals (Li & Lei, 2015). Some techniques are suggested for estimating absolute model fit; posterior predictive model checking (PPMC), limited-information goodness-of-fit statistics, mean absolute difference (MAD), and root mean square error of approximation (RMSEA) are among them. For detailed information on the differences among these techniques and how they are applied, readers are directed to Kunina-Habenicht, Rupp, & Wilhelm (2012), DiBello *et al*. (2007), and Rupp *et al.* (2010). Besides model fit indices, some models such as G-DINA (de la Torre, 2011) provide item-level fit statistics. With item fit, the statistical software programs allocate the best models to individual items.

3.3.1.2. Q-matrix validation

Validating the initially-constructed Q-matrix, which is arguably subjective, is another integrative part of the majority of CDMs. As the Q-matrix plays a significant role in the interpretations one wants to make about the test and test-takers , some indices are proposed by the scholars (e.g. de la Torre & Chiu, 2016 for the G-DINA model; de la Torre & Douglas, 2008 for the DINA model) to empirically identify and replace misspecified entries in the Q-matrix. The modified Q-matrix is also subject to some further modifications by the researcher/field experts since statistical packages by themselves are not capable of finding and modifying misspecifications.

3.3.1.3. Participants' parameters/mastery status

The main goal of all CDMs is providing diagnostic information on test-takers" underlying cognitive abilities in a target domain. So it is obvious that the main output of all CDMs have to be allocated to parameters on participants' knowledge. Not all models have the same way of reporting this output. The G-DINA model for instance, provides latent classes (the skill mastery patterns into which respondents are assigned) to deliver information about mastery status of participants. Mastery probability of each attribute for all individuals is reported in R programming package. The Fusion model, as another example, uses EMstats for examinee mastery statistics by means of Arpeggio software. EMstats produces evaluation statistics on an examinee-by-examinee basis, as well as summary statistics over all examinees (Roussos *et al.*, 2007). Depending on the chosen model, the mastery status is reported.

3.3.1.4. Item parameters

Besides participants' parameters, delivering statistics on the features of items in a test is another target of all CDMs. To take the above mentioned examples again, for the G-DINA model in R programming software, a list of probability of success of each latent class for each item is handed over (Ma & de la Torre, 2016). In the Fusion model, IMstats (item mastery stats) describes how well, on an item-by-item basis and on average over all items, the *Arpeggio*

ISSN 2476-5880

MCMC estimates of examinee mastery of each skill correspond to the actual observed performance of the examinees on each item (Roussos *et al.*, 2007).

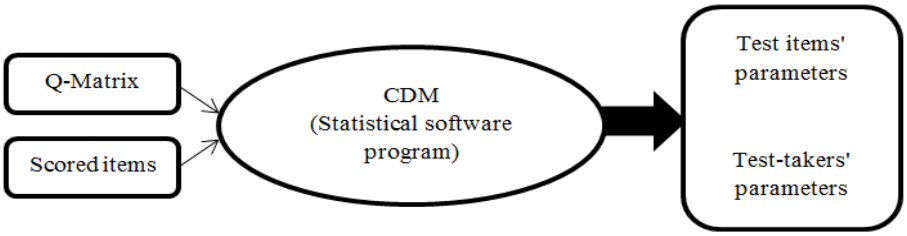

Figure 1. An overall CDA analysis and the outputs

The typical procedure of CDA application, presented in the four main steps, included defining the attributes, developing the Q-matrix, applying a CDM to the Q-matrix and test results, and finally reporting the diagnostic information. To make this procedure more tangible, the procedure of applying a CDM to the data collected from a high stakes test is presented in the next section.

4. A Case Study

In order to illustrate the procedural steps in CDA, we present a case study of applying the G-DINA model on a high-stakes reading comprehension test very briefly. As the sub-skills of reading comprehension tests interact with each other, the selection of the G-DINA model seems logical (for more and detailed information about the model's assumption, refer to de la Torre (2011). The sample test-takers include 2500 randomly selected candidates, both males and females, who sat the university entrance exam in Iran to pursue their MA programs in various fields of Human Sciences in February 2013. The general English reading comprehension section of this exam, which includes 3 passages on different topics and 20 multiple-choice items, was selected for the analysis (the test package number 613 can be retrieved from www.sanjesh.org).

4.1. Extracting attributes

As the study retrofits an existing-non-diagnostically-constructed test, we needed to extract the underlying attributes from the test items. Three main sources were utilized: concurrent literature (Jang, 2005; Gao, 2006; Swaki, Kim & Gentile, 2014; Ravand, 2015; Anani & Javidanmehr (in press)), expert panel's judgment (three PhD candidates of applied linguistics working on educational measurements), and test-takers' think-aloud protocols (10 MA students who had taken the test previously). As the result of this triangulation, four main attributes were defined:

1. *Understanding vocabulary meaning*: Recognizing the meanings of vocabulary items with or without reference to the text.

 ISSN 2476-5880

2. *Making inferences*: Making inferences using explicit information given in the text.

3. *Understanding explicit information*: Recognizing information explicitly mentioned in the text.

4. *Connecting and synthesizing*: Organizing and synthesizing information across parts of the text and understanding the relationship among ideas.

4.2. The Q-matrix: Initial, validated, and final version

The experts decided on the Q-matrix entries (1 for the presence and 0 for the absence of the attributes) first individually and then in a panel discussion session and came to an agreement on the initial Q-matrix. This Q-matrix was validated by the G-DINA package in R programming and the suggested version was provided. As it is suggested by the researchers in the area (e. g., Jang, 2005), this version was scrutinized by the expert panel again to make sure that suggested attributes are truly put in place. The final version of the Q-matrix got prepared, applying experts' suggestions, for the subsequent CDM analysis (Table 3).

Table 3

Initial, Suggested and Final Q-matrices

	Voc	Inf	Exp	Conn		Voc	Inf	Exp	Conn		Voc	Inf	Exp	Conn
41	0	1	1	0	41	0	1	0*	0	41	0	1	**0**	0
42	0	0	1	1	42	0	0	1	1	42	0	0	1	1
43	1	1	0	1	43	1	1	0	1	43	1	1	0	1
44	1	1	0	1	44	1	1	0	1	44	1	1	0	1
45	1	0	1	1	45	1	0	1	1	45	1	0	1	1
46	0	0	1	1	46	0	0	1	1	46	0	0	1	1
47	0	0	1	0	47	1*	1*	1	0	47	**1**	**1**	1	0

 ISSN 2476-5880

48	1	1	0	1	48	1	1	0	1	48	1	1	0	1
49	0	0	1	0	49	1*	1*	1	1*	49	0	**1**	1	**1**
50	1	0	1	0	50	1	1*	1	1*	50	1	0	1	**1**
51	0	1	0	1	51	0	1	0	1	51	0	1	0	1
52	1	1	0	1	52	1	1	0	1	52	1	1	0	1
53	0	0	1	1	53	0	0	1	1	53	0	0	1	1
54	1	0	1	0	54	1	0	1	0	54	1	0	1	0
55	0	0	1	0	55	1*	0	1	1*	55	0	0	1	0
56	0	0	1	0	56	0	0	1	0	56	**1**	0	1	0
57	1	0	0	1	57	1	0	0	1	57	1	0	0	1
58	0	0	1	0	58	0	0	1	0	58	0	0	1	0
59	0	0	1	0	59	1*	0	1	1*	59	0	0	1	0
60	1	0	1	0	60	1	0	1	0	60	1	0	1	0

Initial Q-matrix Suggested Q-matrix Final Q-matrix

*denotes suggested elements highlighted cells are changed

4.3. Data analysis

The data were analyzed in R-programming software, "GDINA" package, version 1.2.1 (Ma & de la Torre, 2017). The estimations in the "GDINA" package are done by MMLE/EM algorithm (de la Torre, 2009; 2011). With the 4 Q-matrix attributes (K=4), 16 latent classes (2^k) were identified. What follows are the data analysis results regarding fit statistics, items and cognitive

attributes' parameters as examples of the analysis. Based on the objectives of each individual study, some other outputs, such as DIF analysis, can also be extracted from the model analysis.

4.3.1. Fit statistics

Absolute fit statistic. As it was discussed earlier, fit statistic is conducted at two levels: *absolute model fit statistic*, which examines the fitness of the model to the data under absolute sense and *relative fit statistic*, which uses a comparative lens to choose the best model from among a bunch of models. The absolute model fit is the prerequisite for subsequent analyses.

Table 4

Item Fit Statistics

	mean[stats]	max[stats]	max[z.stats]	p-value	adj.p-value
Proportion correct	0.00	0.03	**3.60**	0	0.01
Transformed correlation	0.04	0.27	13.34	0	0.00
Log odds ratio	0.27	2.02	14.93	0	0.00

Note: p-value and adj.p-value are associated with max[z.stats].
adj.p-values are based on the bonferroni method.

To consider model-data fit, maximum Z-score of each statistic is evaluated. The rejection of Z-score is the indication of model-data misfit. The significance level of Z-score can be adjusted by means of the Bonferroni correction. For α=0.01, 0.05, and 0.1, critical Z-scores are 4.17, 3.78, and 3.61 respectively. As Table 4 shows, the G-DINA model provide a good test-level fit to the data (Max Z=3.60, α=0.01).

Relative fit statistics. In the G-DINA package different reduced CDMs such as the additive CDM (ACDM; de la Torre, 2011), the reduced reparameterized unified model (RRUM, Hartz, 2002), the DINA, and DINO models can be calibrated as well. In every CDA, the best fitted model needs to be selected for the data analysis. In this regard relative fit statistics, using different indices, illustrate the best model for the specific data and the Q-matrix. In the current study the same procedure was conducted and rival models were compared with the G-DINA model. Table 5 illustrates relative fit indices of conventional Akaike Information Criterion (AIC, Akaike 1973), Bayesian Information Criterion (BIC, Schwarz 1978), and observed log-likelihood ratio test (LRT). The model that reports the least information criteria is considered the

best model to fit the data. As columns 1 and 3 in Table 5 read, G-DINA model with the least log likelihood ratio and AIC indices fits the data significantly better than the other three models. The RRUM comes next on the list. The DINA model though presents the worst model fit.

Table 5

Relative Fit Statistics

Models	#par	logLik	deviance	AIC	BIC	Chisq	Df	p-value
GDINA	119	-14831.34	29662.69	29900.69	30586.19			
DINA	55	-14995.20	29990.41	30100.41	30417.24	327.72	64	<0.001
ACDM	79	-14940.39	29880.78	30038.78	30493.86	218.09	40	<0.001
RRUM	79	-14899.41	29798.81	29956.81	30411.89	136.12	40	<0.001

4.3.2. Attribute prevalence

The estimate of attribute prevalence is provided in Table 6. The indices show both the whole sample's mastery probability of each attribute and the relative difficulty levels of reading comprehension's cognitive sub-skills. As the results show vocabulary knowledge, mastered by 82% of the test-takers , was the easiest attribute and understanding explicit information attribute, mastered by 66% of the tets-takers was considered as the most difficult attribute to master.

Table 6.

The Estimate of Attribute Prevalence

Reading comprehension cognitive attributes	Attribute Prevalence

ISSN 2476-5880

A1. Vocabulary knowledge	0.8276
A2. Making inferences	0.6924
A3. Understanding explicit information	0.6691
A4. Connecting and synthesizing	0.8097

4.3.3. Latent classes' profiles and the posterior probabilities

CDMs categorize test-takers into latent classes, which represent specific mastery/nonmastery profiles for the set of attributes specified in the Q-matrix (von Davier, 2005).Table 7 demonstrates 16 latent classes' profiles and their posterior probabilities. As it is read, the latent class '1111' had the highest class probability (0.340), which means about 34% of the test-takers are expected to have mastered all attribute presented in the Q-matrix. The latent classes '0000' comes second (0.237), which says 23% of the test-takers, were non-masters of all of the attributes. Another latent class which is dominant is '1010', to which 15% of the test-takers belong. This latent class is master of the first and the third defined attributes (vocabulary and explicit information).

Table 7

Latent Classes' Profiles and the Posterior Probabilities

#	Latent class	Posterior probability
1	0000	0.2373
2	1000	0.0373
3	0100	0.0107
4	0010	0.0061
5	0001	0.0428

 ISSN 2476-5880

6	1100	0.0426
7	1010	0.154
8	1001	0.0079
9	0110	0.0476
10	0101	0.1149
11	0011	0.0459
12	1110	0.0301
13	1101	0.0401
14	1011	0.0454
15	0111	0.0461
16	1111	0.3408

4.3.4. Items' probability of success

The probability of answering a certain item correctly having different attribute mastery patterns is another output of the G-DINA model. *P*(1) is the probability of success for those test-takers who have mastered all the required attributes in an item, which involve '1', '11', '111', '1111' in a 4-attribute test. *P*(0), including '0', '00', '000', '0000' in the current test, is the probability of success for those test-takers who have mastered none of the required attributes by means of guessing. *P*(1) of 0.40 or lower indicates difficult items. In this specific test, items 52 to 60 (the last 9 items) seem to be the most difficult ones. *P*(0) above 0.45, on the other hand, indicates answering by guessing. Item 59 (*P*(0)=0.56), which is considered a difficult question as well (*P*(1)=0.0001), is guessed to a large extent.

4.3.5. Item plots

Items' probability of success can also be shown graphically. The plots of items 60 and 49 are shown in Figure 2.

ISSN 2476-5880

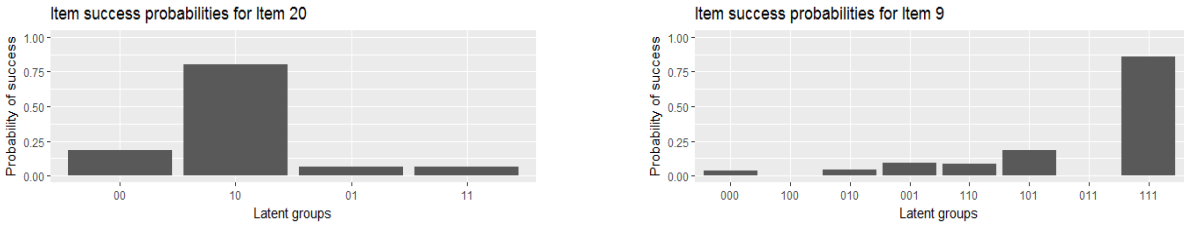

Figure 2. Items' probability of success

The two-attribute item (item 20 in the figure) is a difficult item as the probability of success for the masters of the two attributes (11) is very low (0.06). The probability of getting the item correct for the second class (10), those who were masters of the first attribute is extremely high ($P(10)= 0.80$). The item is not that easy to be guessed by those who have not mastered any of the attributes ($P(00)= 0.18$). On the other hand, the three-attribute item (item 9 in the figure) is not a difficult question as the probability of success for masters of the three presented attributes in getting the item correct is extremely high (0.85). Non-masters (000) do not have the high chance of guessing the item (0.03).

5. Limitations and Drawbacks of CDA

In spite of the advantages and potentials of CDA in advancing educational measurement, some limitations, both theoretical and practical, are in order.

5.1. Subjectivity of attribute definitions and the Q-matrix content

Q-matrix, as the main ingredient of CDA analysis, is constructed based on the defined attributes which are extracted from theories in the literature, test's blueprint, expert panel's judgments, and students' think-aloud; each method alone or in combination. Although the combinations of all methods and statistical Q-matrix validation add to the level of refinement, in the end it cannot be claimed that the results are devoid of subjectivity. As the sources of data in Q-matrix development are human beings, construct under-representation and construct irrelevance variables (Messick, 1995) may threaten the validity of the results. Construct under-representation can be caused by an inappropriate emphasis on particular evidence and arguments and construct irrelevance may define the attributes that are not related to the conceptual framework and the construct in the literature. Accordingly, by misrepresentations of the domain framework, a different Q-matrix with different entries may be handed down for one set of data, and as a result, different outputs are resulted.

5.2. Selection of the most appropriate CDM

Another issue that challenges the CDA application is choosing the most appropriate model from among the different models of analysis for a selected set of data. Some factors such as model-fit evaluation in both absolute and relative terms, sample size, and number of parameters are critical

in data analysis (Torre & Lee, 2013; Li & Lei, 2015). Accordingly, selecting a specific CDM over another might result in various mastery classifications even with the same set of data and participants. Whether the best fitted model, at item level or test level, is selected or not, whether the sample size is large enough to provide accurate results for that specific data, and whether the most appropriate grain size is defined for the number of attributes, various outputs are expected. Highlighting the importance of the best model selection, de la Torre and Lee (2013), for instance, discuss the limitation of their study as even with the same test, set of attributes, and group of examinees, different results may be reported if a different CDM is employed.

5.3. Retrofitting to existing non-diagnostic tests

The result of a CDA study should firstly target the test developers and secondly enhance the process of teaching and learning. Taking into consideration the strengths and weaknesses of the participants and the remedial feedback on the attributes of the target domain, test developers need to review their subsequent tests' blueprints. In this way, the reverse engineering process in CDA is a step forward in educational measurement; however some are worried about the validity of the interpretations. Jang (2008) argues that the non-diagnostic, norm-referenced tests include items that are in line with the psychometric principle essential for creating a bell-shaped score distribution by including a wide range of item difficulty levels. Such a psychometric principle may not conform to the principles that inform diagnostic assessment. Alderson, Brunfaut, and Harding (2014) consider diagnostic assessment as different from achievement tests, placement tests, and proficiency tests as they are designed for purposes other than diagnosis of test-takers' strengths and weaknesses. From their perspective, this process applies ex post facto to tests that have not necessarily been designed with diagnosis in mind.

5.4. Statistical/psychometric knowledge

CDA has been developed in order to enhance educational measurement regarding the fine-grained information it provides, however the level of statistical and psychometric knowledge that it demands impedes its all-inclusive practice. Xie (2016) argues that the need for intensive data processing and computation includes sophisticated psychometric and statistical sub-skills which are not typically available to a university language enhancement program. Every user of CDA needs to know or start learning how to apply a software program linked to a CDM, which is not an easy task. To apply the Fusion model and the G-DINA model, for instance, one needs to run the Arpeggio software and the R-programming package respectively. Not only does the practice needs the statistical knowledge for analysis, but also the interpretation of the CDA jargon in the final reports is not without problems as far as test developers and users are concerned.

5.5. Time issue

Time is another demanding concern. Going through the Q-matrix construction procedure is a very time-consuming task. In both designs, that is to say designing and analyzing cognitive diagnostic tests and reverse engineering existing non-diagnostic tests, the researcher or the users need to triangulate the sources of information to define the attributes, construct the Q-matrix, and

then assign the entries respectively. Sometimes the expectations are far beyond the allocated time frames and the available resources.

5.6. Large samples needed

The data (scored tests) that any CDM statistical package demands is far beyond the classroom scale. Most of the computational tools are designed in such a way that a large sample of participants is needed in order to provide the desirable outputs (Aryadoust, 2011). Large scale assessments are the main targets of cognitive analysis till date.

6. Concluding Remarks

Considering the limitations presented and the further research suggested by scholars in the field, interested researchers are encouraged to go deeper into the following untapped or scarcely tapped areas: First and foremost is the preciseness of the Q-matrix. The validity, reliability, and preciseness of the Q-matrix based on the aforementioned methods need to be examined. Other methods can be designed as well in order to minimize the Q-matrix subjectivity. Introducing cognitive diagnosis assessment into all stages of test construction from the very beginning, instead of only ex post facto CDA design, could be the second mission and a step forward in providing desired feedbacks in educational measurement. Computational statistics could be the third target. Making CDA applicable to almost all educational contexts by means of simpler statistical programs could be addressed as well. How to bring CDA into the borders of classrooms as the most significant assessment context could be tapped by psychometricians and educationists as well. In this way, CDMs can be developed, which do not need that large samples and at the same time present the tangible feedbacks on mastery levels in different sub-skills/attributes for each and every test taker.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, 19*(6), 716-723.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301-320.

Alderson, J. C., Brunfaut, T., & Harding, L. (2014). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*(2), 236-260.

Alderson, J. C. (2010). Cognitive diagnosis and Q-matrices in language assessment": A commentary. *Language Assessment Quarterly, 7*(1), 96-103.

Aryadoust, V. (2011). Application of the fusion model to while-listening performance tests. *Shiken: JALT Testing & Evaluation SIG Newsletter, 15*(2), 2-9.

Blood, I. (2011). Diagnostic second language assessment in the classroom. *Teacher's College, Columbia University Working Papers in TESOL and Applied Linguistics, 11*(1), 57-58.

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.

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement, 70*(6), 902-913.

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

Choi, H-J., Rupp, A. A., & Pan, M. (2012). Standardized diagnostic assessment design and analysis: Key Ideas from Modern Measurement Theory. 61-85.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595-624.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447-468.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31A review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. Volume 26, pp. 979-1030): Elsevier.

Embretson, S. E. (1985). *Test design: Developments in psychology and psychometrics.*

New York: Academic Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380-396.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* University of Illinois at Urbana-Champaign.

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), 191-210.

Jang, E. E. (2008). A Review ofcognitive diagnostic assessment for education: Theory and application. *International Journal of Testing, 8*(3), 290-295.

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

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 31-73.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the test of English as a foreign language (TOEFL).* University of Illinois at Urbana-Champaign.

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.

Lee, Y-W., & Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172-189.

Lee, Y-S., de la Torre, J., & Park, Y. S. (2011). Relationships between cognitive diagnosis, CTT, and IRT indices: an empirical investigation. *Asia Pacific Education Review, 13*(2), 333-345.

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

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.

Li, H., & Suen, HK. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment, 18*(1), 1-25.

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

Li, Hongli. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 17-46.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning -matrix. *Bernoulli (Andover), 19*(5A), 1790-1817.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*(3), 253-275.

McGlohen, M., & Chang, H-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*(3), 808-821.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Mislevy, R. J., Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively Diagnostic Assessment*: Hillsdale, NJ: Erlbaum.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575-603.

Ravand, H., Barati, H., & Widhiarso, W. (2012). Exploring diagnostic capacity of a high stakes reading comprehension test: A pedagogical demonstration. *Iranian Journal of Language Testing, 3*(1), 12-37.

Rojas, G., de la Torre, J., & Olea, J. (2012). Choosing between general and specific cognitive diagnosis models when the sample size is small. *Unpublished manuscript, Rutgers University, New Brunswick, NJ*.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. *Cognitive Diagnostic Assessment for Education: Theory and Applications*, 275-318.

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

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*(2), 461-464.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal, 76*(4), 513-521.

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

Spolsky, B. (1990). Introduction to a colloquium: The scope and form of a theory of second language learning. *TESOL quarterly, 24*(4), 609-616.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. *Diagnostic Monitoring of Skill and Knowledge Acquisition*, 453-488.

Tatsuoka, M. M., & Tatsuoka, K. K. (1989). Rule space. In S. Kotz & N. L. Johnson (Eds). *Encyclopedia of statistical sciences* (pp. 217-220). New York: Wiley.

Torre, J., & Lee, Y-S. (2013). Evaluating the Wald test for item-level comparison of saturated andreduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355-373.

Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics, 14*(2), 60-82.

von Davier, M. (2005). *A general diagnostic model applied to language testing data.* ETS Research

Report RR-05-16. ETS, Princeton, NJ: ETS.

Von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives, 7*(1), 67-74.

Xie, Q. (2016). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward? *Educational Psychology*, 1-22.

Yamamoto, K., & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. *Test Theory for a New Generation of Tests*, 275-295.