Kuramsal Eğitimbilim Dergisi, 11(2), 244-260, Nisan 2018
Journal of Theoretical Educational Science, 11(2), 244-260, April 2018
[Online]: http://dergipark.gov.tr/akukeg
DOI number: http://dx.doi.org/10.30831/akukeg.363915

Where the Rivers Merge: Cognitive Diagnostic Approaches to Educational Assessment

Eğitimde Ölçme ve Değerlendirme Uygulamalarına Bilişsel Tanılayıcı Bir Yaklaşım

Tuğba Elif TOPRAK [*] **Abdulvahit ÇAKIR**[**]

Received: 08 December 2017 **Accepted:** 08 March 2018

ABSTRACT: A growing emphasis on the union of cognitive psychology with psychometrics has led to the inception of Cognitive Diagnostic Assessment (CDA). CDA can be defined as a cognitively-grounded assessment methodology which aims to detect examinees' strengths and weaknesses in a given domain, make reliable diagnostic classifications directly from the statistical models, and present stakeholders with fine-grained and pedagogically-meaningful diagnostic feedback. Although CDA holds great promise for educational assessment practices, it remains relatively unknown to many assessment specialists. Hence, this paper aims to describe the theoretical underpinnings and working principles of CDA by giving information about the developments that have led to the inception of CDA and elaborate on how CDA can be implemented in operational assessment settings either by using an inductive or retrofitted approach to foster learning and enhance accountability within educational programs. Finally, the potential that CDA bears for educational assessment is discussed and practical implications are made.

Keywords: cognitive diagnostic assessment, diagnostic classification modelling, skills diagnosis, assessment for learning.

ÖZ: Bilişsel psikolojinin psikometri ile harmanlanması Bilişsel Tanılayıcı Değerlendirme (BTD) adı verilen ölçme ve değerlendirme yaklaşımının ortaya çıkmasını sağlamıştır. BTD, bilişsel temelli, istatistiki açıdan sofistike ve alternatif bir ölçme ve değerlendirme yaklaşımıdır. Bireylerin belirli bir beceri ya da akademik alandaki güçlü ve zayıf yanlarının, eksiklerinin ve yanılgılarının saptanmasını ve bu hususlara yönelik paydaşlara (öğrenci, öğretmen, veli ve idarecilere) bireylerin halihazırdaki durumları hakkında detaylı dönüt verilmesini amaçlar. Sağlanan dönüt, pedagojik açıdan anlamlı ve öğrenme sürecini destekleyici boyutta olmalıdır. Bu değerlendirme yaklaşımının eğitim öğretim faaliyetleri için pek çok yararı olmasına karşın, BTD hem eğitim araştırmacıları hem de ölçme değerlendirme alanında çalışan araştırmacılar tarafından yeteri derecede tanınmamaktadır. Bu makalede, BTD yaklaşımının ortaya çıkmasına sebep veren eğitimsel akım ve gelişmeler ele alınmış, BTD'nin kuramsal temelleri, çalışma prensipleri, işlevleri hakkında detaylı bilgi verilmiştir. Ayrıca, BTD'nin öğrenme çıktılarını iyileştirme ve eğitim programlarının kalite ve hesap verebilirliğinin artırılması hedeflerine yönelik olarak, eğitim ve ölçme değerlendirme ortamlarında nasıl uygulanabileceği hususunda öneriler sunulmuştur.

Anahtar kelimeler: eğitimde ölçme ve değerlendirme, bilişsel tanılayıcı değerlendirme, tanılayıcı sınıflama modellemesi.

[*] *Corresponding Author*: PhD, Gazi University, Ankara, Turkey, tetoprak@gazi.edu.tr
[**] Prof. Dr., Gazi University, Ankara, Turkey, vahit@gazi.edu.tr

Citation Information
Toprak, T. E., & Çakır, A. (2018). Where the rivers merge: Cognitive diagnostic approaches to educational assessment. *Kuramsal Eğitimbilim Dergisi [Journal of Theoretical Educational Science], 11*(2), 244-260.

ISSN: 1308-1659

Introduction

Cognitive Diagnostic Assessment (CDA) aims to help furnish fine-grained diagnostic feedback about individual test takers' mastery of a set of skills in a given domain by classifying test takers into skill mastery classes and reporting their mastery profiles in great detail. CDA approach blends theories of cognition of interest with statistically sophisticated measurement models and pinpoints to individual test takers' cognitive strengths and weaknesses in a defined domain or skill (Jang, 2008; Rupp & Templin, 2011). Besides classifying test takers into the skill mastery classes and diagnosing their current status, CDA fulfils several valuable functions such as monitoring the diagnostic quality of test items, evaluating the effectiveness of estimation process and examining he cognitive processes and mechanisms that are essential to the successful execution of test tasks. In particular, CDA approach has been motivated and facilitated by the recent developments in the fields of cognitive psychology and psychometrics. Thus, CDA could be regarded as a confluence where the theories of cognitive psychology and psychometrics at the macro level, and the theories of educational assessment and domain of interest at the micro level, merge.

Since Snow and Lohman's call (1989) to incorporate the principles of cognitive psychology into educational assessment, the union between these two fields has been on the agenda of many assessment researchers (Embretson, 1991; Embretson & Gorin, 2001; Gierl, 2007; Mislevy, 1996; National Research Council (NRC), 2001; Roussos, DiBello, Stout, Hartz, Henson, & Templin 2007). This call may have partly been motivated by the predominance of large-scale testing and the extensive practice of unidimensional IRT (Item Response Theory) based assessments in most educational settings. IRT-based large-scale assessments, which are usually high-stakes and unidimensional in nature, have benefited assessment practices with increasing reliability and accountability (Roussos et al., 2007). However, these assessments have mostly concentrated on rank-ordering test takers along a continuum in a given domain, skill or on a global ability, and mainly served selection, placement and admission purposes (Roussos, Templin, & Henson, 2007). The inevitability of large-scale assessments in many areas is evident, but their limitations have also been a matter of discussion in the field of educational assessment (Anastasi, 1967; Lee & Sawaki, 2009a; Leighton & Gierl, 2007b; Nichols, 1994; Snow & Lohman, 1989).

Particularly, a few of these limitations are worth mentioning. One significant limitation may be the dearth of fine-grained and pedagogically meaningful information that could be extracted from such assessments. Large-scale assessments could yield information about test takers' global ability in a given domain and provide stakeholders (i.e. test takers, teachers, parents, and institutions) with a quick snapshot of test takers' current status, yet this picture would fall short in revealing the details about such status. Another limitation could be linked to the granularity of the construct of interest since large-scale assessments tend to focus on general abilities such as math and language ability, often providing a rather broad construct definition. This limitation runs the risks of construct under-representation and presenting stakeholders with a single test score, which usually remains as a coarse indicator of test takers' overall ability. Thus, discussions centring on these limitations have underlined the urgency of a reconceptualization in testing theory, and consequently, assessment researchers began to ponder on issues such as the function and capacity of large-scale assessments as pedagogically informative tools, the alignments between the characteristics of the constructs, how these constructs are defined in operational large-scale testing situations, and the validity of interpretations arising from such assessments (Huff & Goodman, 2007).

One notable effort that was made as a response to these limitations would be the report entitled "Knowing What Students Know", released by the NRC in 2001. The NRC report defined assessments as evidentiary systems that are comprised of three different components specified as cognition, observation, and interpretation. According to this conceptualization, cognition stands for the assumption that ability is cognitively-grounded, latent and unobservable, while observation refers to the process of data collection by using a specifically designed tool for assessment purposes. The last component, interpretation, requires drawing inferences about test takers' ability or knowledge based on the observations made. In fact, more than a decade before the NRC report was released, a similar argument was put forward by Messick (1989), an influential figure in the history of educational assessment. In his seminal chapter on validity, Messick (1989) argued that understanding cognitive processes underlying the test performances would add to the construct validity of a test, and maintained that advances in cognitive psychology could contribute to this understanding considerably.

The union between cognitive psychology, learning sciences and educational assessment, in addition to helping gain a better understanding of the construct at hand, is expected to yield several beneficial outcomes; inter alia, coming up with viable ways of assessing that construct, modelling the cognitive mechanisms and processes required for working on an assessment task and generating more insightful and exploratory theories of learning and teaching (Snow & Lohman, 1989). Since traditional criteria such as item difficulty, item discrimination and examining test specifications would not be adequate alone to meet these demands, educational assessment researchers have begun to seek novel ways and methods that may prove more beneficial and feasible in gathering further information regarding the functionality, validity, and reliability of test items and tasks, and more importantly, information about the test takers themselves (Rupp, 2007). To this aim, one strand of research focused on examining cognitive processes needed for solving test items and modelling item statistics (e.g., Carr, 2003; Embretson, 1998; Freedle & Kostin, 1993; Kostin, 2004), while another strand of research included scale anchoring studies (e.g., Beaton & Allen, 1992; Gomez, Noah, Schedl, Wright, & Yolkut, 2007; Liao, 2010) and factor analytic studies (e.g., Davis, 1944; Spearitt, 1972; Thorndike, 1971) to better understand the properties of test items. However, these applications were limited in that they examined the group level performances rather than individual test taker performances, fell short in embracing the current cognitive theories (Gao & Rogers, 2011), and failed particularly in obtaining detailed information about test taker profiles (Lee & Sawaki, 2009a). Consequently, the shortcomings of these methods in yielding meaningful information about test takers' performances and the underlying traits leading to those performances and the increasing need and for fine-grained and diagnostic feedback have altogether led to the inception of an alternative approach to educational assessment; namely Cognitive Diagnostic Assessment (CDA). CDA is a relatively new, albeit flourishing field with a firm ground in several seminal works; such as Embretson's (1983) paper on blending cognitive psychology with construct validation, Messick's chapter on validity (1989), Snow and Lohman's (1989) chapter on combining cognitive psychology with educational measurement, Nichols' paper in which he coined the term 'cognitive diagnostic assessment' (1994), and the coedited books on CDA by Nichols, Chipman and Brennan (1995), Leighton and Gierl (2007a), and Rupp, Templin and Henson (2010). Alongside this work addressing mainly the theoretical underpinnings of CDA, there is a growing body of empirical research exploring the potential of CDA in operational settings as well (e.g., Buck, Tatsuoka, & Kostin, 1997; Chen, Ferron, Thompson, Gorin, & Tatsuoka, 2010; Im & Park, 2010; Jang, 2005; Jurich & Bradshaw, 2013; Kim, 2015; Lee & Sawaki, 2009b; Sawaki, Kim, & Gentile, 2009). Moreover, in order to render CDA applicable to practical testing situations, numerous multidimensional measurement models a.k.a., diagnostic classification models (DCMs) with different statistical assumptions and functions have also been generated and presented to the use of CDA researchers (e.g., de la Torre & Douglas, 2004; Embretson, 1984; Gierl, Cui, & Hunka, 2008; Hartz, 2002; Junker & Sijtsma, 2001; Mislevy, Steinberg, &Almond, 2003; Templin & Henson, 2006; von Davier, 2007). More information on these up-and-coming models is provided in the following sections.

What is Cognitive Diagnostic Assessment?

To better communicate what CDA is, it may be helpful to decompose the label into the two adjectives it features; cognitive and diagnostic. The term cognitive, to begin with, is used in a different sense in CDA than it is used in the fields of cognitive science and computer sciences. CDA employs the term cognitive to refer to the assessments that rely on cognitive models. At this point, further explanation is clearly called for another term, cognitive models, to understand in what sense CDA uses the term. In CDA literature, cognitive models are defined as test theories that are useful for diagnosing cognitive mechanisms (Nichols, 1994), and generated by scrutinizing the skills, knowledge, and processes that are used by test takers while working on a test task (Gierl & Cui, 2008). Put differently; cognitive models help assessment specialists explain and predict test takers' performances. CDA employs cognitive models so as to i) generate items tapping on specific skills, knowledge and/or cognitive processes which are called attributes in CDA literature; ii) depict item-attribute alignments for the existing tests, and iii) make interpretations about test takers' performances on test tasks (Gierl, Cui, & Zhou, 2009). By utilizing cognitive models, CDA is assumed to be capable of capturing the existing deficiencies or gaps in the cognitive mechanisms and knowledge structures that are of great importance to perform on a test task. Cognitive models are created by reviewing the theoretical literature in the area of interest and backed up by empirical research that investigates which knowledge structures and cognitive processes are inherent or crucial to the successful execution of the construct at hand.

In CDA literature; skills, knowledge structures, and processes that test takers should possess to get an item correctly are called attributes (Buck & Tatsuoka, 1998; Gierl, Leighton, & Hunka, 2000) and the relationships between the items and attributes are expressed in an incidence matrix called the Q-matrix (Tatsuoka, 1983). In most CDA applications the Q-matrix has typically been treated as a cognitive model guiding the analyses and shaping the interpretations arising from the assessment (See Table 1 for a hypothetical Q-matrix for a second language reading comprehension test). A Q matrix is an item by attribute indicator which shows the attributes that need to be mastered to answer each item correctly and it indicates the relationships between the items and attributes through the numbers 1 and 0. To illustrate, based on the hypothetical Q-matrix presented in Table 1, it can be deduced that Item 1 measures only the third attribute while Item 2 taps all the attributes listed. Item 7, on the other hand, measures the first and second attributes, but not the third one. The construction of the Q-matrix is of utmost importance to CDA since its completeness and robustness exerts a drastic impact on the results that the application yields and the validity of the interpretations drawn from these results (Madison & Bradshaw, 2014; Rupp & Templin, 2008). To construct and refine the Q matrix, several methods and tools such as think-aloud protocols, content analysis, and expert panels may be employed. Hence, it would be fair to say that one of the features that distinguishes CDA from many classroom-based and large-scale assessments would be the inclusion of a cognitive model which forms the basis of the construct definition and test design and contributes significantly to the cognitive focus of the assessment.

Table 1

A Hypothetical Q-matrix for a Second Language Reading Comprehension Test

	Attribute 1	*Attribute 2*	*Attribute 3*
Item	Understanding explicitly stated information	Inferencing	Connecting and synthesizing
1	0	0	1
2	1	1	1
3	1	1	0
4	1	0	1
5	0	1	0
6	0	1	1
7	1	1	0
8	0	1	0
9	0	0	1
10	1	1	1

The second adjective, diagnostic, on the other hand, communicates a functional quality. The diagnostic focus of CDA, which specifically aims to obtain rich, fine-grained and detailed information about test takers' mastery status and performance in a given field, could be another important quality that distinguishes CDA from other forms of assessments. From a practical perspective, it is possible to conceive that virtually all classroom-based and large-scale assessments may yield diagnostic information about the current status of test takers. For instance, an English language teacher, after giving language learners a test on grammar, although roughly, could spot the areas that a language learner might struggle while dealing with English grammar.

Likewise, the results of a proficiency or a placement test of English held at a higher education institution could provide stakeholders with a broad picture of the test takers' performance in second language listening. However, this picture would remain over-simplified and low-on capturing details when compared to a picture that is taken through the lenses of CDA. Extracting some diagnostic feedback from every kind of assessment might make sense at first sight; yet, there is a pitfall in viewing non-diagnostic assessments as being equal to assessments that are designed solely for diagnosis and classification purposes.

Apart from diagnosis and classification purposes, CDA may serve a great number of purposes such as; i) assessing each test taker based on a level of competence in a set of skills ii) providing stakeholders with fine-grained diagnostic feedback pertaining to the test takers' abilities that are under scrutiny, iii) evaluating the diagnostic quality of individual test items and the test itself, iv) evaluating the effectiveness of the estimation process, v) enabling to gain deeper insights into the nature of the construct underlying the test performances, vi) enhancing the construct validity and increasing the reliability of the classifications, and vii) peeking inside the cognitive processes and mechanisms that the test takers are likely to engage with while working on a test task. To illustrate, a notable example depicting how CDA was effectively applied and how these purposes were achieved to a considerable degree would be Jang (2005), in which CDA was applied to the reading section of the Next Generation TOEFL IBT. Jang (2005) was able to estimate each test taker's mastery profile of second language reading comprehension, present test takers' with a detailed diagnostic report on their ability, determine how attributes interacted with each other, assess the diagnostic quality of the test items, evaluate the effectiveness of the estimation, and gain an understanding of the cognitive processes and mechanisms that test takers may use with while responding to the test items.

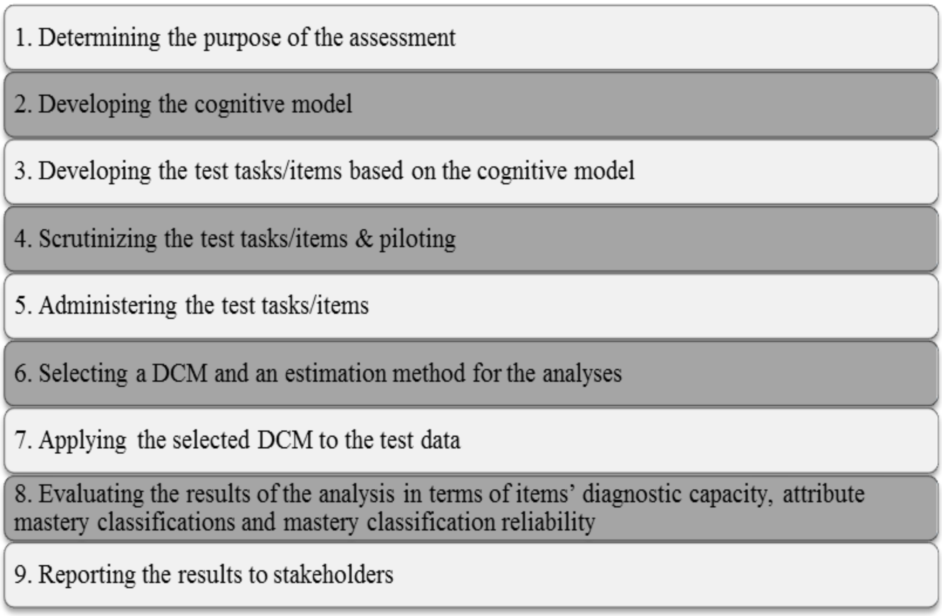
As indicated earlier, CDA is primarily designed to assess each test taker on a set of skills to obtain rich and detailed diagnostic information about the test takers' mastery status, weaknesses, and strengths in a given domain. Hence, perhaps, the most useful implication that CDA mainly holds for teaching and learning contexts would be how it views and treats assessment. In contrast to traditional approaches to educational assessment, CDA views and treats assessment as a tool for facilitating learning and makes a distinction between assessment of learning and assessment for learning by favouring the latter (Jang, 2008, 2009b). This approach may prove more useful since it empowers teachers by providing them with fine-grained and diagnostic information that would help modify and tailor their teaching according to the current status and needs of learners. CDA aims to evaluate how much a test taker knows about a subject, how well s/he performs in an area of interest and how much remains to be acquired. Furthermore, it sheds light on which processing skills that form the basis of successful performance are attained, and which remain yet to be acquired; which knowledge structures are missing from the cognitive base, and which misconceptions exist to block the successful execution of basic knowledge structures and processing skills. For no doubts, working towards these goals would indicate moving beyond the traditional way of assigning a single score to each test taker, which usually stands as a coarse indicator of test takers' current status.

In addition to increasing the granularity of the construct being measured and the volume of feedback provided, CDA may prove more beneficial than traditional approaches to assessment especially in terms of item design, for it enables test developers to check the diagnostic quality of test items in greater depth (Yang & Embretson, 2007). Traditional approaches to item design primarily target creating items for stable latent traits. In such an approach, item quality is generally maintained by checking the relationships among the items themselves and other external traits, and by meeting the psychometric standards such as item facility and discrimination. The relationships between the items, constructs and the domain of interest are usually expressed in rather general terms, such as in the form of guidelines or table of specifications. Moreover, traditional item design usually does not involve an understanding of how the item properties influence the cognitive processes triggered by the items or which cognitive processes are required to get these items right. On the contrary, CDA incorporates the notion of substantive theory into the test design before embarking on item construction. The substantive theory, appertaining to the area of interest, elaborates on the mental processes and mechanisms that are likely to affect the performance and forms the backbone of cognitive models. If the substantive theory is specified in advance and the items are constructed based on this very theory, besides strengthening the construct validity, it would also be possible to detect the gaps in knowledge structures, deficiencies in skills, and misconceptions about phenomena.

Thus, CDA entails posing a disciplined and well-informed approach to the item construction and test design, an endeavour which needs to be backed up with the substantive theory that is made up of the current theories of cognitive psychology and learning sciences, as well as the theories of the domain of interest. For instance, while deciding about which skills and processes to include in a second language reading comprehension test, CDA draws on the current theories and body of empirical research on second language reading comprehension in areas such as cognitive psychology, second language literacy, second language education and language assessment. Moreover, since CDA requires test developers to come up with a well-informed and well-articulated construct theory, it could be regarded as a strong program of validity as well (Leighton & Gierl, 2007a). In addition to the substantive theory, overarching assessment frameworks such as Evidence-Centered Design (Mislevy, Steiberg, & Almond, 2003) and Cognitive Design System (Embretson, 1998) could also be incorporated into CDA. These frameworks may prove useful in guiding assessment specialists along with a series of assessment phases such as defining the construct of interest, creating test tasks, administration and validation of these tasks. When it comes to the application of CDA, a disciplined and well-informed approach, and a series of steps, each feeding and informing the subsequent steps are needed. Overall, in CDA framework, the first step is determining the purpose of assessment and defining learning and instructional goals that serve as the criteria for the diagnosis. Then, test tasks (e.g., test items) are created and scrutinized before selecting an appropriate psychometric model, which is referred to as a diagnostic classification model (DCM) for the sake of consistency throughout the paper. After appropriate statistical methods that would be used for estimation are determined, analyses are carried out. Finally, the results of the analyses are evaluated and reported to the stakeholders. So far, there have been two approaches to the application of CDA; these are inductive and retrofitted or posthoc analysis approaches (Roussos et al., 2007). The inductive approach entails developing a cognitive model and a diagnostic test from ground-up. Then, the test data obtained from the administration of the diagnostic test are analysed through a DCM. This approach allows for capturing the characteristics of test items in detail. The retrofitted or posthoc analysis approach, on the other hand, uses an already existing test in the hope of extracting useful diagnostic information. In this approach, the results of a non-diagnostic test are analysed by using a DCM. Due to its convenience, the majority of CDA applications have so far been in the form of retrofitting, in which the data from high-stakes tests such as TOEFL (Test of English as a Foreign Language), IELTS (International English Language Testing System) and SAT (Scholastic Aptitude Test) have been analysed through a DCM. Given these large-scale and high-stakes assessments which are primarily generated for accountability and rank-ordering purposes are indispensable in many educational settings, the principles of CDA can help extracting more fine-grained and diagnostic information from these assessments (Huff & Goodman, 2007). Furthermore, in countries such as the USA, England and Germany where performance standards are applied nationwide to maintain accountability, feedback systems which would not only provide information about the current status of learners but also point to potential remedial pathways for problem areas are needed (Rupp & Mislevy, 2007).

Although previous retrofitting applications have yielded beneficial results to some extent, a great deal of studies have also highlighted several potential drawbacks with the retrofitted approach and argued that the inductive approach would produce more promising results when compared to its retrofitted counterpart (Gierl & Cui 2008; Jang, 2009; Kim, 2015; Lee & Sawaki, 2009b; Rupp & Templin, 2008). Figure 1 below depicts how CDA could be implemented using the inductive approach. It should be noted that except the first five steps, the retrofitted approach follows the same steps as in the inductive approach. Since the former reverse-engineers existing tests, the first five steps are taken in considerably different ways across the two approaches.

Figure 1. CDA Overall Application Process



1. Determining the purpose of the assessment

2. Developing the cognitive model

3. Developing the test tasks/items based on the cognitive model

4. Scrutinizing the test tasks/items & piloting

5. Administering the test tasks/items

6. Selecting a DCM and an estimation method for the analyses

7. Applying the selected DCM to the test data

8. Evaluating the results of the analysis in terms of items' diagnostic capacity, attribute mastery classifications and mastery classification reliability

9. Reporting the results to stakeholders

So far, the basics and theoretical underpinnings of CDA approach have been elaborated on. The next section delves into more detail about CDA's statistical background and introduces its functional agents, which are called diagnostic classification models (DCMs).

How does CDA work?

CDA is not basically interested in a single score, or how many items have successfully been answered by a test taker, but it is more concerned with understanding the response patterns that involve different cognitive processes, skills, and knowledge structures helping answer the items correctly (Yang & Embretson, 2007). To illustrate the logic behind CDA and provide a synopsis of how it works before going into the details, let us think about a scenario in which three individuals take the same test of general language ability in English. The scores obtained by Student A, B, and C, are 70, 75 and 82 respectively. In such a case, traditional testing procedures would rank these test takers along a continuum, and conventional interpretations arising from such assessment would indicate how a test taker performs when compared to the other test takers in the group. Although such assessments would be feasible in cases where a selection or admission is made, assigning each test taker a single score would certainly not be adequate if the chief purpose is to make diagnostic decisions. Specifically focusing on making diagnostic classifications, CDA framework breaks the construct of interest into a set of attributes to increase the granularity of the diagnostic feedback. CDA, thinking back to our scenario, focuses on a narrower ability such as second language reading comprehension in the language arts, or adding/subtracting skills in mathematics, and divides these abilities into smaller components called attributes. Let us think that for second language reading comprehension ability, these attributes are determined to be inferencing, understanding explicitly stated information, summarizing and, understanding grammar and sentence structure. Although the underlying second language reading comprehension ability may be continuous, these attributes are assumed to be categorical in nature, and they divide test takers into two groups; as masters and non-masters. After the relationships between the items and attributes are expressed in the Q-matrix, DCMs place test takers into these mastery groups by tracking test takers' response patterns to the items measuring the attributes.

Before the application of the DCMs to make diagnostic classifications, initially, a cognitive model establishing the link between the substantive theory and assessment design is constructed. Next, test items are created either from ground-up or through reverse-engineering, and the alignments between the items and attributes are expressed in the form of a Q-Matrix. Then, DCMs are executed to analyse the test data and estimate mastery profiles of the test takers. A mastery profile, which is denoted as αi and defined as a vector of length K, (K refers to the total number of attributes) shows which attributes are mastered and which are not. The mastery profiles estimated through DCM applications help researchers make a diagnostic classification (Henson & Douglas, 2005). Thus, it would not be an exaggeration to say that DCMs may be considered as the backbone and main agent of the CDA applications, and to understand how CDA works; one needs to grasp what DCMs are and how they function. DCMs have been given different names in relevant literature such as cognitive psychometric models (Rupp, 2007), restricted latent class models (Haertel, 1989), structured located latent class models (Xu & von Davier, 2008), cognitive diagnosis models (Nichols et al., 1995) and latent response models (Maris, 1995). Although each naming indicates a certain characteristic of these psychometric models such as their purpose or statistical properties, regardless of what label is used, these models help define a test taker's ability in a given domain based on the attributes that have or have not been mastered. This mastery profile allows for determining the probability of a correct response for each item, and DCMs can effectively predict the probability of each test taker's falling into a specific latent diagnostic latent class (Henson, Templin, & Willse, 2009).

DCMs are statistically sophisticated multidimensional measurement models and have some features in common with other measurement models such as Classical Test Theory (CTT), Item Response Theory (IRT), FA Factor Analysis (FA) and Latent Class Analysis (LCA). A comprehensive definition of DCMs has been offered by Rupp and Templin (2008) in which they described DCMs as:

> probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and non-compensatory ways to generate latent classes" (p. 226).

Moving on from this definition, one could say that DCMs are probabilistic models since they make information about the probabilistic attribute profile of each test taker available, indicating whether they have mastered one or more than one attribute and estimate the probability of each test taker's being a member of a specific latent diagnostic class. DCMs are multidimensional, for they are capable of partitioning the underlying ability into a set of subskills and classifying the test takers into the latent mastery classes based on their mastery or non-mastery of specified attributes, unlike unidimensional IRT models which usually rank-order test takers along a continuum of a single general ability (Madison & Bradshaw, 2014). Furthermore, DCMs, when compared to other multidimensional IRT models, are also reported to display an increased reliability and feasibility even with fewer items, bringing researchers a significant advantage in operational testing settings (Bradshaw, Izsak, Templin, & Jacobson, 2014). DCMs are confirmatory in that they carry out the analyses based on the Q-matrix which displays the relationships between the latent variable that is being measured and the observable variables, i.e. test items. The Q-matrix shows which skills, strategies or attributes are needed to answer each question correctly and specifies the loading structure of DCMs (Li & Suen, 2013; Sawaki et al., 2009). This way, CDA resembles Confirmatory Factor Analysis (CFA). However, DCMs deal with modelling categorical rather than continuous latent variables. That is one of the reasons DCMs could serve better for making diagnostic classifications and decisions when compared to other measurement models since using categorical latent variables is more efficient for making a classification (Templin & Bradshaw, 2013). Moreover, DCMs can also be utilized to test researchers' hypotheses about the cognitive processes that test takers are assumed to engage with while working on test tasks. Hence, DCMs may function as a tool for scrutinizing researchers' theory-based conjectures, and collecting empirical evidence which helps researchers shape and refine their construct definitions (Bradshaw et al., 2014).

There are more than 60 DCMs listed in the relevant literature (Fu & Li, 2007) and it might be a challenging task for researchers to select, use and optimize a particular

DCM. Although the underlying idea behind CDA applications is similar across all DCMs, these statistical models may show substantial variations in the ways they define the probability of a correct response. These variations in statistical modelling stem from the differences induced by choice of a cognitive theory which depicts how the skills aggregate to lead to an item response behaviour (Rupp & Templin, 2008). In other words, the theory of how cognitive processes and skills impact an item response behaviour may shape the decisions that researchers make in opting for a particular DCM. Basically, concerning item-attribute relationships, DCMs are classified into two categories; non-compensatory and compensatory models. In the first category, the non-compensatory models, a deficit in one attribute cannot be compensated for by a surplus in another attribute. Put differently, the conditional relationship between any attribute and the item responses depends on the remaining required attributes that have been mastered or not. Due to this dependency, non-compensatory models are further divided into two groups; conjunctive and disjunctive models. A typical example of conjunctive models would be the Deterministic Input; Noisy “And” Gate (DINA) Model. DINA is very restrictive in that the probability of a correct response is only high when the test taker masters all the attributes required for an item (Haertel, 1989; Junker & Sijtsma, 2001). Disjunctive models, on the other hand, as in the Deterministic Input; Noisy “Or” Gate (DINO) model (Templin & Henson, 2006), assume that the mastery of additional attributes leads to an increase in the probability of a correct response once a subset of the required attributes has been mastered. In contrast to the non-compensatory models, the second category, compensatory models, posit that the mastery of a skill can compensate for the non-mastery of other skills. A characteristic example of compensatory models would be the compensatory Reduced Unified Model (RUM) (Hartz, 2002) which postulates that a deficit in one attribute can be made up by a surplus in another attribute.

To date, especially, conjunctive models have mostly been utilized in fields such as mathematics where the task of interest can be broken down into its smaller units, and successful completion of the task depends on the completion of each unit. However, recently, the use of compensatory DCMs has increasingly gained popularity when compared to their conjunctive counterparts and have been applied to measure relatively more complicated constructs in fields such as the language arts, where skills may function in a highly interactive and compensatory fashion (Stanovich, 1980) and consequently, a surplus in one skill may compensate for the lack of another skill.

Conclusions and Implications

Shifting sands and changing winds in the field of educational assessment apparently call for more cognitively-grounded, substantively backed-up and diagnostic assessment designs. Such assessments are believed to provide stakeholders with more pedagogically-meaningful, fine-grained and individualized feedback that could benefit them in many ways. For instance, by obtaining such feedback, educators may be able to tailor their teaching practices and check whether their instructional decisions lead to desirable outcomes; test takers may be able to become more aware of their strengths and weaknesses in an ability, and resultantly grow to be more responsible for their learning; parents may be more willing and able to collaborate with students and their teachers to help them achieve their goals; and finally institutions may track and assess their educational quality and policies more effectively.

In contrast to the traditional approaches to assessment that focus on rank-ordering test takers in a given domain, CDA, which pinpoints to the deficiencies, misconceptions, and weaknesses of test takers in a given domain, also has more penetration. In educational contexts, penetration can be defined as the quality of extracting more detailed and rich information from test scores which help shed light on cognitive processing, knowledge structures and concepts that test takers possess (Gorin, 2007). Moreover, the value of the contribution that CDA may offer to educational assessment can easily be deduced from the perspective that it holds on assessment; that is bringing the assessment of learning rather than the assessment of learning into sharp focus (Jang, 2009). In addition to functioning as an assessment methodology that has great implications for learning and teaching, CDA also helps reinvigorate the validity of interpretations arising from diagnostic assessments with producing ample statistical evidence. Nevertheless, there are several problems that need to be sorted out to maximize the true potential of CDA. These problems can briefly be cited as lacking truly cognitive diagnostic tests that have been created solely for diagnosis and classification purposes, the issue of rendering the results of the CDA applications more comprehensible to the stakeholders, the need for utilizing complementary data collection tools to increase the validity CDA results and lacking essential software and expertise to carry out CDA.

First, the majority of CDA applications have been in the form of retrofitting, and if carefully designed and conducted, these assessments also do have the potential to yield effective results. The relevant literature points to the need for more assessments that are solely designed for CDA since in applications where a DCM is retrofitted to non-diagnostic assessments, problems related to model fit, item characteristics, test takers' mastery classifications and model convergence problems may be expected (Rupp & Templin, 2008). At this point, although it would require a more time-consuming and laborious process of creating a diagnostic test from scratch and gathering a large amount of test data to meet the statistical requirements of DCMs, the inductive approach would prove more beneficial than its retrofitted counterpart. To this end, while taking an inductive approach to CDA and designing tests for diagnostic purposes from ground-up, the diagnostic quality and capacity of items can be enhanced by creating test items tapping common misconceptions held in the domain of interest. Moreover, the distractors can be arranged in such a way to address different levels of misconceptions that are common among test takers and knowledge states that are likely to be missing or poorly-constructed.

Second, the issue of diagnostic reporting does not seem to receive the attention it deserves. While a number of studies have been concerned with the application of DCMs in operational settings, research exploring the impact of CDA applications on teaching and learning contexts remains relatively limited. Clearly, more research efforts are needed to explore the ways to better communicate the results of CDA applications to stakeholders, understand how educators and test takers make sense of CDA results and to what extent they incorporate the interpretations arising from these results in their educational practices, and capture far-reaching and long-term effects on CDA on educational contexts. For no doubts, making the most of CDA relies on incorporating assessment results into classroom practices effectively. Hence, the issue of how these results are reported and disseminated to stakeholders grows increasingly significant. Results of the CDA need to be interpretable by stakeholders so that they would guide and aid educators in planning, executing, assessing and tailoring their educational practices.

Third, we should caution that CDA does not do the magic on its own and requires a well-informed approach to the test design that is backed up with the substantive theory and complementary data collection tools. Conclusions drawn from CDAs can be supported, validated and enriched by using various tools and techniques such as think-aloud protocols, eye-tracking methodology, interviews, content analysis and expert panels. These tools and techniques have extensively been used during Q-matrix construction process to ensure the completeness and robustness of the Q-matrix. Nonetheless, their use does not need to be limited to the Q-matrix construction and validating the interpretations arising from these assessments, but could extend its scope to the making of cognitive models and test task generation. Thus, considerable attention and effort should be invested in CDA applications which are of inductive nature and backed up with an array of data collection tools facilitating the application of CDA and ensuring the triangulation of CDA results.

Fourth, the lack of user-friendly and free available software to apply CDA, and the lack of expertise among educational researchers to handle the complexity of the DCMs may pose serious challenges to the researchers who would like to apply CDA. Unfortunately, although the number of DCMs available exceeds 60 (Fu & Li, 2007), there is not much software that is user-friendly, publicly available and practical to calibrate CDA data, and this situation seems to hamper the application of CDA in many areas greatly. Furthermore, from a methodological point of view, CDA draws on from the advancements in psychometrics and applied cognitive psychology; and from a practical point of view, it aims to apply and transfer these advancements to the assessments in fields like language arts, science, and mathematics. Thus, continued and closer collaboration between the experts and practitioners in these target fields and psychometricians is highly needed to ensure the match between theory and practice. We hope that addressing these issues would spur much methodological and theoretical advancement in CDA research.

References

Anastasi, A. (1967). Psychology, psychologists, and psychological testing. *American Psychologist, 22*, 297-306.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*, 191-204.

Bradshaw, L., Izsak, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33,* 2-14.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*(3), 423-466.

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.

Carr, N. T. (2003). *An investigation into the structure of text characteristics and reader abilities in a test of second language reading* (Unpublished PhD dissertation). University of California, Department of Applied Linguistics, Los Angeles, USA.

Chen, Y. H., Ferron, J. M., Thompson, M. S., Gorin, J. S., & Tatsuoka, K. K. (2010). Group comparisons of mathematics performance from a cognitive diagnostic perspective. *Educational Research and Evaluation*, *16*(4), 325-343.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69,* 333-353.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175-186.

Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 37*, 359-74.

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

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

Freedle, R., & Kostin, I. (1993). *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items.* (TOEFL Research Reports No. RR-93-44). Princeton, NJ: Educational Testing Service.

Fu, J., & Li, Y. (2007, April). *Cognitively diagnostic psychometric models: An integrative review*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Gao, L., & Rogers, T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing, 28*(1), 77-104.

Gierl, M. J. (2007). Attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement 44*(4), 325-340.

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

Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*(3), 293-313.

Gierl, M. J., Leighton, J. P., & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule space model for test development and analysis. *Educational Measurement: Issues and Practice, 19*, 34-44.

Gomez, P. G., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, *24*(3), 417-444.

Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 173-201). New York: Cambridge University Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

Hartz, S. M. (2002). *A Bayesian guide for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois, Department of Statistics, Urbana-Champaign, IL.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*, 262-277.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika, 74,* 191-210.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). New York: Cambridge University Press.

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: linkage to instruction. *Educational Research and Evaluation*, *16*(3), 287-301.

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

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y. R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University.

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

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

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

Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing, 14*(1), 37-41.

Kim, A. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 3*(2) 227-258.

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items* (TOEFL Research Rep. No. RR-79). Princeton, NJ: Educational Testing Service.

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

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

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

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

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.

Madison, M., & Bradshaw, L. (2014). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491-511.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

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

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-62.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington DC: National Academy.

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

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT based latent class models. *Journal of Educational Measurement, 44*, 293-311.

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

Rupp, A. A. (2007): The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing, 7*(2), 95-125.

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive Foundations of Structured Item Response Models. In J. P. Leighton & M. J. Gierl (Eds.*), Cognitive diagnostic assessment in education: Theory and applications* (pp. 205-341). New York: Cambridge University Press.

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

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

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.

Snow, R. E., & Lohman, D. F. (1989). Implication of cognitive psychology for education measurement. In R.L. Linn (Ed.*), Educational measurement* (pp. 263-331). New York: Macmillan.

Stanovich, K. E. (1980). Towards an interactive compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*, 32-71.

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

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*, 251-275.

von Davier, M. (2007). *Hierarchical general diagnostic models* (Research Report No. 07-19). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008). *Linking for the general diagnostic model*. ETS Research Report. Princeton, New Jersey: ETS.

Yang, X., & Embretson, S. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 119-145). New York: Cambridge University Press.