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

Diagnostic Test Construction: Insights from Cognitive Diagnostic Modeling

Somaye Ketabi[1]*, Dr. Seyyed Mohammed Alavi[2], Dr. Hamdollah Ravand[3]

Received: 30 December 2020 Accepted: 1 February 2021

Abstract

Although Diagnostic Classification Models (DCMs) were introduced to education system decades ago, it seems that these models were not employed for the original aims upon which they had been designed. Using DCMs has been mostly common in analyzing large-scale non-diagnostic tests and these models have been rarely used in developing Cognitive Diagnostic Assessment (CDA) from scratch. Despite the prevalence of *retrofitting* CDA studies, *true* applications of CDA are believed to be rare since, firstly, a coherent framework to conduct such studies had not been available and, secondly, researchers were not able to analyze various DCMs according to the same model fit indices and criteria. This paper presents a summary of different types of DCMs and reviews *true* and *retrofitting* CDA studies. Having examined the limitations of previous CDA studies, the present study argues for the implication and application of Ravand and Baghaei's (2019) framework to conduct *true* CDA studies. This framework is of importance since not only does it fit into prominent frameworks in education assessment such as Cognitive Design System and Assessment Triangle, but also it can provide test-developers with practical steps in conducting valid cognitive diagnostic tests.

Keywords: Cognitive Diagnostic Assessment; Diagnostic Classification Models; Q-Matrix construction; *retrofitting* CDA study; *true* CDA study

1. Introduction

Diagnostic Classification Models (DCMs), as Rupp and Tumplin (2008) believe, can estimate test-takers' chances of answering an item correctly according to the attributes that each item measures. Employing a DCM requires some steps to be taken one of which is constructing a Q-matrix according to the opinions of the experts or test-takers. The Q-matrix specifies which attributes are (going to be) measured by each item (Tatsuoka, 1983). It should be mentioned that DCMs are psychometric models in which test tasks are decomposed into the processes needed to complete each task (Whitely, 1983). Since interpretations based on DCMs are discrete and criterion-referenced, provided that a valid Q-matrix is developed, DCMs can be used in diagnostic assessments (Tatsuoka, 1983).

Ravand and Baghaei (2019) classified CDA studies into three main categories. First, in *true* CDA studies, researchers design tests with diagnostic purposes from the beginning and develop the items according to a constructed Q-matrix, and the results obtained can be used to

[1] English Department, Faculty of Foreign Languages and Literatures, University of Tehran, Iran
[2] English Department, Faculty of Foreign Languages and Literatures, University of Tehran, Iran
[3] English Department, Faculty of Literature and Humanities, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

provide valuable information about test-takers' strengths and weaknesses. In the second category, *retrofitting* studies, existing non-diagnostic high stakes tests are used to extract information about the test and test-takers, and no items are actually constructed. In the last group of studies, researchers analyze the technical aspects such as model selection and fit to build infra structure for the application of different DCMs (Ravand & Baghaei, 2019).

Retrofitting CDA studies have been popular among researchers and most of them analyzed various non-diagnostic high-stakes tests using DCMs (e.g. Chen & Chen, 2016; Jang, 2005; Javidanmehr & Anani Sarab, 2019; Kim, 2015; Lee & Sawaki, 2009b; Li & Suen, 2013; Mirzaei, Vincheh, & Hashemian, 2020; Yi, 2012). Although *retrofitting* studies may shed light on the technical issues of selecting DCMs, Gorin (2009) and Tatsuoka (2009) argue that applying DCMs to tests which were originally unidimensional would lead to serious consequences. The original theories based on which such tests were constructed were IRT and Classical True Score (CTS), and these theories are applicable when an item measures one attribute only. However, in analyzing large-scale tests and constructing Q-matrix, it is possible that several attributes may be assigned to an item which was unidimensional according to its underlying theory. The ramification of the dimensionality conundrum creating by *retrofitting* is reflected in (1) highly correlated dimensions when DCMs are retrofitted to assessments designed through IRT and CTS (Ravand & Baghaei, 2019), (2) low scale reliability for the individual dimensions since there may not be sufficient number of items measuring each attribute (Kunina-Habenicht, Rupp, & Wilhelm, 2017), and (3) poor model fit results. With the above points in mind, caution should be exercised to practice *retrofitting* only when it is not possible to design a test from the scratch within a diagnostic framework. It should be mentioned that even the few *true* CDA studies (e.g., Liu et al., 2013; Ranjbaran & Alavi, 2017) have chosen the DCMs arbitrarily regardless of the degree of match between the assumptions of the models and how the attributes underlying the test are assumed to interact. Choice of the DCMs has mostly been driven by software availability and familiarity in these studies. Arbitrary choice of DCMs might result in misleading information regarding the classification of the test takers, which is the prime objective of CDA studies.

No matter what type of CDA study is going to be conducted, steps taken must be based on a framework and each selection must be justified by a plausible reason. Ravand and Baghaei (2019) suggested a set of steps to apply DCMs in *true* or *retrofitting* studies. This paper presents a summary of major DCMs and then reviews different types of CDA studies which were conducted and mentions their major limitations. Then, the researchers describe how Ravand and Baghaei's (2019) framework is in line with two of the most significant assessment frameworks, how it can be applied in *true* CDA studies, and what challenges researchers may face in implementing such a framework. Strategies which can be used in meeting these challenges are also suggested in different parts.

2. Major Diagnostic Classification Models

A major categorization of DCMs classifies them as being compensatory (disjunctive) versus non-compensatory (conjunctive) (Roussos, Templin, & Henson, 2007). In compensatory models, high competence in an attribute of a skill can compensate for the lack of competence in another attribute in completing a task; however, mastery over more attributes cannot increase

the chance of completing the task successfully. Deterministic Input, Noisy "Or" Gate Model and Noisy Input, Deterministic "Or" Gate Model (DINO & NIDO; Templin & Henson, 2006), are two examples of compensatory/disjunctive models. In a non-compensatory model, on the other hand, non-mastery of one attribute cannot be compensated by mastery over other attributes. As Li et al. (2015) claimed, non-compensatory/conjunctive models, e.g. Deterministic Input, Noisy "And" Gate Model (DINA; Junker & Sijtsma, 2001) and Noisy Inputs, Deterministic "And" Gate (NIDA; DiBello, Stout, & Roussos, 1995) have been more common in conducting cognitive diagnostic analysis. The popularity of non-compensatory models might be due to the fact that the first applications of DCMs were mostly in mathematics where all predetermined stages should be gone through in order to answer a math question and competence in one attribute cannot compensate for non-competence in another (Roussos et al., 2007).

In a more recent categorization, additive models allow for more flexibility in the relationship between attributes such that the subset of required attributes mastered by a test-taker affects the probability of answering an item correctly (Ravand & Baghaei, 2019; Chiu, Koehn & Wu, 2016). Reduced Reparameterized Unified Model (RRUM; Chiu, Koehn & Wu, 2016), Additive CDM (ACDM; de la Torre, 2011), Compensatory Reparameterized Unified Model (C-RUM; DiBello et al., 1995; Hartz, 2002), Non-Compensatory Reparameterized Unified Model (NC-RUM; Hartz, 2002), and Linear Logistic Model (LLM; Maris, 1999) are some examples of additive models.

de la Torre's (2008) application of Hierarchical DINA (HO-DINA) Model and Templin and Bradshaw's (2013) use of Hierarchical Diagnostic Classification Model (HDCM) created a new extension of specific and general models; i.e. hierarchical models, in which structural relationships among attributes of a skill are considered. In other words, hierarchical models are able to capture the impact of sequential order of teaching materials which is needed in testing those materials taught in a specific order in some instructional syllabi.

In another categorization, DCMs are classified as being saturated or constrained (Li, Hunter, & Lei, 2015) or as Ravand (2016) mentioned, they are general or specific. In a saturated or general model, different kinds of relationships among compensatory, non-compensatory, or additive attributes can be investigated. The General Diagnostic Model (GDM; von Davier, 2005), Log-Linear Cognitive Diagnostic Model (LCDM; Henson, Templin, & Willse, 2009), and Generalized Deterministic-Input, Noisy "And" Gate Model (GDINA; de la Torre, 2011) are three examples of saturated models. Although it is reasonable to employ a saturated model because of its flexibility, constrained or specific models are less complex and can be used with a smaller sample size compared to saturated models. It should be mentioned that as Rojas, de la Torre, and Olea (2012) argued, using simpler models such as constrained models can provide more meaningful results.

3. CDA Studies and their Limitations

Among various *retrofitting* studies, the following were significant since they applied different DCMs in large-scale tests. Lee and Sawaki (2009b) investigated different DCMs using the performance of test-takers in listening and reading sections of iBT TOEFL and compared the classifications of the NC-RUM (Fusion Model), Latent Class Analysis Model (LCA;

Yamamoto, 1990), and the GDM. They found similar categorization of these model and argued that the NC-RUM, the LCA Model, and the GDM could accurately classify test-takers according to their mastery or non-mastery of different attributes. Yet, firstly, Lee and Sawaki (2009b) did not report values of model fit indices in their analysis, secondly, their main comparison was according to the univariate and bivariate distributions of individual attributes, and thirdly, they did not compare models of different types (compensatory and non-compensatory) and retrofitted the existing data from a largescale non-diagnostic test instead of conducting a *true* CDA study.

Yi (2012) used the data of Lee and Sawaki (2009b) in a *retrofitting* study and compared model fit indices and univariate and bivariate distributions of individual attributes in a general model (LCDM), an additive model (C-RUM), a non-compensatory model (DINA), and two compensatory models (DINO and NIDO), and found that the C-RUM classified test-takers like the LCDM. However, Yi (2012) did not compare models in terms of consistency and accuracy in classifications.

Li et al. (2015) used the data of the reading section of the Michigan English Language Assessment Battery and compared the fit indices of the DINA, DINO, ACDM, NC-RUM, and G-DINA. Like Yi (2012), they found that the additive model under study, i.e. the ACDM, fit the data like the general model, i.e. the GDINA, and could classify models more accurately than others. Just like previous studies, Li et al. (2015) did not report values of classification consistency and accuracy and did not conduct a *true* CDA study.

In a comprehensive study, Ravand and Robitzsch (2018) analyzed the performance of a general model (GDINA), three additive models (ACDM, C-RUM, NC-RUM), a compensatory model (DINO), and a non-compensatory model (DINA) by employing the data of a high-stakes reading comprehension test and found that the GDINA had the best values in terms of model fit indices and classification consistency and accuracy, and additive models (C-RUM, NC-RUM, and ACDM) showed close resemblance to the general model. They also compared models at item level and concluded that some models might fit the data at this level because the relationship between reading attributes would be a combination of compensatory and non-compensatory ones. Ravand and Robitzsch's (2018) study can be of importance due to the variety of comparison criteria employed not only at test level but also at item level. However, just like previous studies, these researchers did not compare models in a *true* CDA study and retrofitted existing data.

Among researchers who tried to conduct a *true* CDA study, Ranjbaran and Alavi (2017) developed a reading comprehension test based on CDA but only used one pre-determined CDM, i.e. RUM, for their analysis. As Ravand and Baghaei (2019) explained, researchers mostly choose the CDM according to the availability of the necessary software programs and their knowledge to use various software applications. However, relationships among reading attributes within a reading comprehension test are more complex than the underlying assumptions of a single DCM, and various DCMs should be compared according to their statistics and parameters to choose the most appropriate model. Moreover, Ranjbaran and Alavi (2017) refined the Q-matrix according to the results obtained in Think-aloud Protocol. Yet, in a *true* CDA study, the Q-matrix is a fixed factor and the items must be revised according to the Q-matrix.

In another study in which the authors claimed to have developed a test based on CDA, Liu et al. (2013) constructed a computerized adaptive English achievement test with cognitive diagnostic approach. The researchers tried a post-hoc approach in which a Q-matrix was developed after choosing the items. They analyzed the items by employing only one DCM, i.e. the DINA model. It should be mentioned that the main purpose of Liu et al. (2013) was to use the CDA to provide more comprehensive feedback for the students, their teachers, and their parents. But their approach in employing CDA was not a *true* one. Using the DINA model may be useful in providing informative reports of students' strengths and weaknesses in each attribute, but this model may not fit the data of such a study since it was only designed to analyze non-compensatory relationship among different attributes.

4. A Framework to Conduct *True* CDA Studies

Ravand and Baghaei (2019) developed two separate frameworks to conduct *true* and *retrofitting* studies which can be applied to tests assessing different constructs. Nevertheless, construct related to productive skills, i.e. writing and speaking, may require Q-matrix construction procedures which are different from those of mathematics and reading comprehension. Figure 1 represents the steps needed to conduct a *true* CDA study according to Ravand and Baghaei's (2019) framework. The steps suggested by Ravand and Baghei (2019)P are in line with the properties of Cognitive Design System (Embretson, 1994) and the principles of the Assessment Triangle (National Research Council, 2001).

The main properties of the conceptual and procedural frameworks of the Cognitive Design System (Embretson, 1994) are followed by Ravand and Baghaei (2019) steps to conduct a *true* CDA study. As for the first characteristic, Embretson (1994) prioritized explicit procedures in preparing the test content. The practical steps suggested by Ravand and Baghaei (2019) including consulting expert judgement and employing Think-aloud protocol are some methods which can be used in preparing the test content. Employing Think-aloud Protocol is itself a means to create a link between the meaning of scores and underlying cognitive processes which is the second property of the Cognitive Design System (Embretson, 1994). Ravand and Baghaei (2019) named Think-aloud Protocol as a useful method in which examinees reflect on the underlying processes at the time of taking the test. The third property of the Cognitive Design System as listed by Embretson (1994) is the representation of cognitive complexity by the item parameters. Ravand and Baghaei (2019) observed this principle in the procedures they suggested in constructing the Q-matrix and mentioned that the complexity of an item could be controlled by the nature and number of its measured attributes. Embretson (1994) highlighted the underlying cognitive processes in doing a task in the last property of the Cognitive Design System again and emphasized the link between these processes which define the task difficulty and the skill or ability. This link is made by the several methods (e.g., expert judgment and a review of related theories) suggested by Ravand and Baghaei (2019) and the attributes, as the building blocks of a skill, which require more complex processes would result in more difficult items or tasks.

According to the National Research Council (2001), Assessment Triangle has three pillars which are cognition, observation, and interpretation. Cognition is the theory of the structure of the skill and how it is acquired, observation includes the techniques employed to

 ISSN 2476-5880

collect evidence about examinees' proficiency in the skill, and interpretation is the statistical model used to draw inference from the observed performance. An ideal assessment is the one in which each of these components are related to each other; i.e., the cognitive theory is the basis of the observation method, and the statistical model examines whether the observed performance is related to the theory of the skill. The steps suggested by Ravand and Baghaei (2019) in their framework precisely aim at making the link between the theory, observation, and interpretation. Reviewing the related literature, consulting expert judgment, and using examinees' comments through Think-aloud interviews indicate whether the items actually measure the attributes based on which they were constructed. Moreover, the method of selecting the CDM according to various fit indices and indicators emphasizes the importance of interpreting test scores and choosing the correct statistical model in Ravand and Baghaei's (2019) framework.

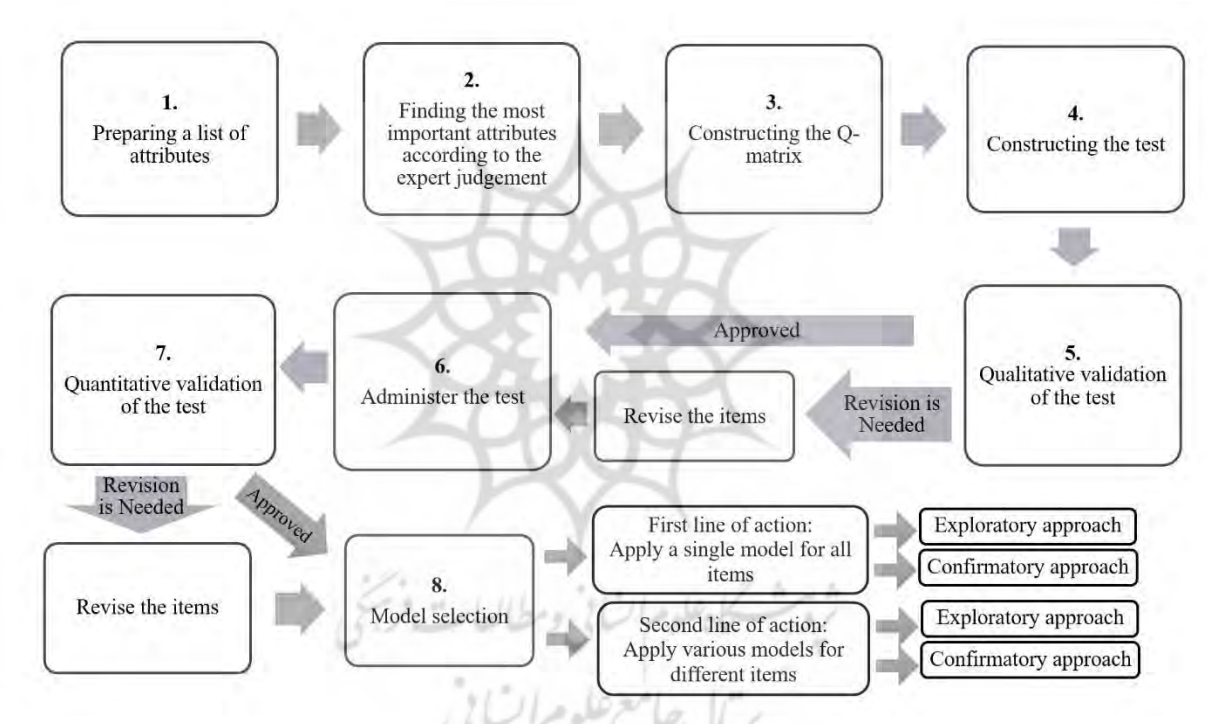

Figure 1. Framework Proposed to conduct a *true* CDA study (Ravand and Baghaei 2019)

4.1.Preparing a List of Attributes

The first step in Ravand and Baghaei's (2019) framework is to review relevant theories of the construct under study and to use expert judgment and findings of previous *true* and *retrofitting* CDA studies to prepare a list of attributes. This step is taken in order to find a cognitive theory of test performance, and as Ravand and Baghaei (2019) mentioned, theories of this type are rare and researchers should employ an implicit theory of test performance by consulting related sources and literature. As another valuable source not mentioned by Ravand and Baghaei (2019) is the course books usually studied by the students to improve the subskills or attributes related to the construct under study. A survey can be conducted in order to find the most frequently-used course books and the addressed attributes therein can be included in the list.

ISSN 2476-5880

4.2.Finding the Most Important Attributes

The second step in Ravand and Baghaei's (2019) framework is to ask at least five expert judges to rate the importance of the attributes included in the list. Numbers can be used in order to indicate the significance of various attributes (e.g. a scale of 1 to 5 in which 1 means the least important and 5 means the most important) and those attributes rated at least by two thirds of the judges must be considered for Q-matrix construction. Judges at this stage must have some important characteristics. First, they must be familiar with the CDA and its major objectives. Second, they must have sufficient knowledge of the target participants and what attributes they have to acquire in order to be successful in their future uses of the second or foreign language.

4.3.Constructing the Q-Matrix

As Gorin (2009) mentioned, the most challenging and important stage in a CDA study is constructing the Q-matrix since the validity and accuracy of claims made about test-takers' mastery or non-mastery over attributes depend upon the accuracy of the Q-matrix. Three points must be considered in developing a Q-matrix: 1) Q-matrix specification, i.e. attributes measured in each item should be specified precisely; 2) Q-matrix design, i.e. arrangement of attributes in the Q-matrix; and 3) grain size of each attribute, i.e. level of specification in each attribute (Ravand & Baghaei, 2019).

Rupp and Templin (2008) found that lack of correct specification of attributes in the Q-matrix would result in slipping or guessing parameters. Slipping parameter, or incorrect answer to an item by a participant who has mastered all its attributes, occurs if attributes are deleted from a Q-matrix incorrectly, and guessing parameter, or correct answer to an item by a participant who has not mastered all its attributes, would be seen if attributes are added to a Q-matrix incorrectly. Therefore, adding or deleting attributes may result in misclassification and considering the goal of CDA studies which is classification of test-takers, researchers must carefully consider all variables which would have an effect on the accuracy of classifications.

Classification accuracy can also be affected by the design of the Q-matrix as well. For instance, De Carlo (2012) showed that in the DINA model each attribute should be measured in isolation at least once. Moreover, Madison and Bradshaw (2015) found that attributes assessed in isolation were measured more correctly than those in combination in the LCDM. They noted that if two attributes were always measured together, they should be combined and considered as a single attribute. But a very important factor to bear in mind is that when test-takers answer some items, they usually employ more than a single attribute and may refer to multiple attributes to answer them correctly. Therefore, in assigning attributes to different items, researchers must proceed with caution so that not many uni-attribute or multi-attribute items are included.

The number of attributes or their grain size is also important in developing a Q-matrix. As Ravand and Baghaei (2019) mentioned, the more specified attributes are, the more informative they will be for instructors. However, large number of attributes will make interpretations difficult (Xu & Zhang, 2016). Diagnostically, a skill should be specifically defined in terms of its attributes no matter how many attributes there are, but the number of latent classes will increase dramatically when more attributes are added to a DCM. In other words, if there are n attributes, the number of latent classes will be 2n. The more latent classes

are, the more items and participants are required. de la Torre and Minchen (2014) specified that the number of attributes should not exceed 10. Yet, as Ravand and Baghaei (2019) explained, 10 attributes result in 210=1024 latent classes, and with 1000 participants, there will be 1000/1024=0.98 participants in each class. To put it another way, with 1024 latent classes and 1000 test takers some latent classes would be either assigned to a small number of test takers or no one at all.

As another important point, Ravand and Baghaei (2019) mentioned that the level of test-takers' language proficiency is also important in coding items. For instance, if test-takers are at elementary stage of acquiring basic grammar and vocabulary, items should be coded for these attributes. Yet, when test-takers have already acquired basic vocabulary and syntax, using these two attributes in coding items would result in low discrimination, and test-takers cannot be classified as masters and non-masters correctly according to these attributes.

4.4.Constructing the Test

In order to construct the test, three issues, i.e., item types, subject of materials, and the difficulty level of materials and items must be observed. First, regarding the type of the items, for example, it should be decided whether they are going to be just multiple-choice items or gap-filling ones. This selection partly depends on the feasibility of checking and coding each type of items. Multiple-choice items can be a good option in assessing receptive skills by considering the large number of participants needed in a CDA study and the feasibility of checking these items. Second, the subject of materials used in tests plays an important role. For instance, if test-takers are already familiar with "Fluid Mechanics", choosing a passage with a content related to this area may harm the validity of such a test because a group of test-takers may answer the items correctly due to their familiarity with the subject and another group may not succeed in doing so because they lack that specific knowledge. It is suggested that researchers select authentic materials not specific to a field of study to avoid such problems. Third, the difficulty level of the material and the items is another important factor. As an example, if the passages of a reading test are too easy to understand or the questions in a speaking test are too simple, test-takers will not use various resources to answer them and the result will not be reliable for a CDA study in which test-takers' strengths and weaknesses in different attributes are in focus. Some techniques such as comparing the readability of passages in the test with those in frequently-used course books and asking expert judgement can be employed to choose materials with appropriate difficulty level.

4.5.Qualitative Validation of the Test and Revising the Items

In order to check whether the items actually assess the attributes specified in the Q-matrix, researchers can use two qualitative methods. First, they can give the test and a list of attributes to a group of experts and ask them to choose the attributes each item measures. As another method, researchers can employ Think-aloud Protocol. The group of test-takers participating in Think-aloud protocol should be familiar with different subskills of the construct under study so they will be able to express their underlying thinking processes better than participants who are not familiar with the construct and its attributes or are not competent enough. As to the think-aloud Protocol, the retrospective method is preferred over concurrent one because in the

latter test-takers' proficiency in the skill or their verbalization of the underlying processes may be hampered due to the fact that participants must focus on two tasks of answering questions and verbalizing their thoughts simultaneously; therefore, they may have problems in performing both tasks as perfectly as possible. In retrospective Think-aloud Protocol, on the other hand, there is a risk of forgetting the flow of thoughts, thus researchers should try to keep the time lapse between taking the test and participating in the interview session as minimum as possible. One solution to keep this lapse short is using several interviewers. For instance, if ten participants are going to take part in Think-aloud Protocol interview, they can be divided into two groups (five participants each) and each group can take the same test in two consecutive days and participate in a Retrospective Think-aloud interview afterwards (after a short break). Five interviewers who are informed about the procedure of Think-aloud Protocol can help the researcher. It is better to present the participants with their own exam papers and ask them to talk about their thinking process in their mother tongue since as Mackey and Gass (2015) claimed, asking students to verbalize their thoughts in a second or foreign language may hinder their natural flow of thoughts.

4.6.Administrating the Test

When choosing the participants, researchers must be careful in selecting those whose level of proficiency over the skill was in focus in test construction. Due to the large number of participants needed in CDA studies, if it is not possible to administer the test to all intended test-takers in a single session, researchers can do so in several sessions.

4.7.Quantitative Validation of the Test

Ravand and Baghaei (2019) proposed a quantitative approach after administrating the test through the procedure suggested by de la Torre and Chiu (2016) to ensure that the relationship between the attributes and the items does actually exist. de la Torre and Chiu (2016) introduced a discrimination matrix which can indicate misspecified items in the Q-matrix. Using this method, de la Torre and Chiu (2016) were also able to suggest a new pattern of attribute-item relationship in a Q-matrix. Although employing this quantitative method can result in a valid Q-matrix in *retrofitting* CDA studies, researchers may face a considerable challenge in using de la Torre and Chiu's (2016) suggested procedure in *true* CDA studies. A large number of participants are needed to extract this discrimination factor, and researchers may need to do the validation procedure several times. For instance, de la Torre and Chiu (2016) used the data of 536 participants. The obtained result would suggest a new Q-matrix but in a *true* CDA study the Q-matrix is fixed and the items must be revised to match the original Q-matrix. One solution is revising the items and checking them several times (with the help of many participants each time) until the suggested Q-matrix matches the original one. This method will definitely require a lot of time and energy.

4.8.Model Selection

In their framework, Ravand and Baghaei (2019) argued that two lines of action can be followed by researchers. In the first one, which is more popular, the researcher uses a single model for all the items whereas in the second one the researcher applies various models for different

individual items. These two lines of action can be followed by using an exploratory or confirmatory approach. The first approach is exploratory in which the researcher does not have obvious theoretical reasons to apply a single pre-determined model and opt to use several models either at the test or item level and then rely on statistics and choose the best model according to model fit indices. In the confirmatory approach, however, the researcher chooses a single model which matches the insights driven from the relevant theories. Therefore, according to Ravand and Baghaei's (2019) framework, there are four possibilities in model selection procedure. Obviously, the first two concern test level applications and the next two are related to item-level applications.

a) In the first line of action within the confirmatory approach, researchers choose a model based on some theoretical evidence, apply it to the whole test and then check the model fit indices.
b) In the first line of action within the exploratory approach, researchers apply different types of models across the board and then compare the model fit indices.
c) In the second line of action and confirmatory approach, different models are selected for different items, the multi-DCM model is run, and model fit indices are checked.
d) In the second line of action and exploratory approach, researchers run the G-DINA model and individual items choose their own model. Model fit indices of different models are also checked.

5. Conclusion

The main purpose of the present overview was to explain the first framework to conduct a *true* CDA study, the challenges which researchers may face in employing it, and possible solutions to overcome them. The impact and importance of this framework cannot be denied since one of the main reasons behind the scarcity of *true* CDA studies was lack of a coherent set of procedures. Another reason might be the various indices produced by each software application to do the statistical analyses of different models and researchers' inability to compare different indices and select the best model accordingly. Considering the fact that multiple models can be compared in R software by using the same model fit indices, researchers are now able to design *true* cognitive diagnostic tests. Although the framework suggested by Ravand and Baghaei (2019) explains each step to conduct a *true* CDA study in detail and R software has already helped researchers to run the required analyses, the framework is still in its infancy and may need some revisions when used in different settings. Last but not least, the major factor in a cognitive diagnostic assessment is the feedback which has not been addressed sufficiently in the literature. The main purpose of this kind of assessment is to find and then improve the weak points in language learners and this part of CDA can be investigated in future studies.

Declaration of Conflicting Interests

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

References

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

Chiu, C. Y., Köhn, H. F., & Wu, H. M. (2016). Fitting the reduced RUM with Mplus: A tutorial. *International Journal of Testing, 16*(4), 331-351. doi: 10.1080/15305058.2016.1148038

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

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

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

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. doi: 10.1007/ s11336-008-9063-2

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología educativa, 20*(2), 89–97. doi: 10.1016/j.pse.2014.11.00

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

Embretson, S. (1994). Applications of cognitive design systems to test development. In *Cognitive assessment* (pp. 107-135). Springer, Boston, MA. doi: 10.1007/978-1-4757-9730-5_6

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

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

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*, 191-210. doi: 10.1007/s11336-008-9089-5

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

Javidanmehr, Z., & Anani Sarab, M. R. (2019). Retrofitting Non-diagnostic Reading Comprehension Assessment: Application of the G-DINA Model to a High Stakes Reading Comprehension Test. *Language Assessment Quarterly, 16*(3), 294-311. doi: 10.1080/15434303.2019.1654479

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272. doi: 10.1177/01466210122032064

Kim, A.-Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258. doi: 10.1177/0265532214558457

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2017). Incremental validity of multidimensional proficiency scores from diagnostic classification models: An illustration for elementary school mathematics. *International Journal of Testing, 17*(4), 277-301. doi: 10.1080/15305058.2017.1291517

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. doi: 10.1080/15434300903079562

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. doi: 10.1080/10627197.2013.761522

Li, H., Hunter, C. V., & Lei, P. W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33*(3), 391-409. doi: 10.1177/0265532215590848

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

Ma, W. & de la Torre, J. (2018). GDINA: The generalized DINA model framework. R package version 2.7.9. Retrieved from https://CRAN.R-project.org/package=GDINA

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design*. Routledge.

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

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

Mirzaei, A., Vincheh, M. H., & Hashemian, M. (2020). Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context. *Studies in Educational Evaluation, 64*, 100817. doi: 10.1016/j.stueduc.2019.100817

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. In J. Pelligrino, N. Chudowsky, & R. Glaser (Eds.), *Board on testing and assessment, center for education. Division of behavioral and social sciences and education.* Washington, DC: National Academy Press.

Ranjbaran, F., & Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation, 55*, 167-179. doi: 10.1016/j.stueduc.2017.10.007

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

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24-56. doi: 10.1080/15305058.2019.1588278

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

Robitzsch, A., Kiefer, T., George, A., C., & Uenlue, A. (2018). CDM: Cognitive diagnosis modeling. R package version 7.4-19. Retrived from https://CRAN.R-project.org/package=CDM

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

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills Diagnosis Using IRT- Based Latent Class Models. *Journal of Educational Measurement, 44*(4), 293-311. doi: 10.1111/j.1745-3984.2007.00040.x

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*, 219-262. doi: 10.1080/15366360802490866

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

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. Routledge. doi: 10.4324/9780203883372

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. doi: 10.1037/1082-989X.11.3.287

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251–275. doi: 10.1007/s00357-013-9129-4

von Davier, M. (2005). *A general diagnostic model applied to language testing data (RR-05-16).* Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2005.tb01993.x

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

Xu, G. and Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*(3), 625–649 doi: 10.1007/s11336-015-9471-z

Yamamoto, K. (1990). HYBILm: A computer program to estimate HYBRID model parameters. *Princeton, NJ: Educational Testing Service*. doi: 10.1002/j.2333-8504.1995.tb01637.x

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