Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 11, April 2015 ISSN 1531-7714

Cognitive Diagnostic Modeling Using R

Hamdollah Ravand, *Vali-e-Asr University of Rafsanjan, Iran*
Alexander Robitzsch, *Federal Institute for Education Research, Innovation & Development of the Austrian School*

Cognitive diagnostic models (CDM) have been around for more than a decade but their application is far from widespread for mainly two reasons: (1) CDMs are novel, as compared to traditional IRT models. Consequently, many researchers lack familiarity with them and their properties, and (2) Software programs doing CDMs have been expensive and not readily available. The present paper presents a reader-friendly introduction to the CDMs and uses the CDM package (Robitzsch, Kiefer, Cathrice George, & Uenlue, 2014) in R to demonstrate the application of the generalized deterministic-input, noisy-and-gate model (G-DINA; de la Torre, 2011) and interpret the output. R is a free open source tool which can be utilized to conduct a range of CDMs which otherwise would need separate software programs to run each.

Cognitive diagnostic models (CDMs) are receiving increasingly more attention in conferences, journals, and books. They have the capability to provide detailed diagnostic feedback about the reason why a given test taker might succeed or fail on any given test. Although researchers and practitioners are getting more and more aware of the CDMs and their effectiveness in personifying the "assessment *for* learning rather than assessment *of* learning" motto, CDMs have remained underutilized for two major reasons (de la Torre, 2009): (a) As compared to traditional IRT models, CDMs are relatively novel and in some cases, more complex. Consequently, many researchers lack familiarity with these models and their properties (b) Unlike traditional IRT models, which can be analyzed using commercially available software, accessible computer programs for CDMs are not readily available.

In what follows, a brief introduction of CDMs is presented. Then a discussion of the advantages of using the CDM package in R is in order. Furthermore, data from Ravand (in press) will be used to walk the readers through the R code and the steps required to conduct CDM and an accessible annotated presentation of outputs generated by the CDM package is provided. The data for the present study were a random sample (n =5000) of the applicants into the English master programs at state-run universities in Iran. University Entrance Examination for Master programs at state universities (UEE) is a high-stakes test that screens the applicants into English Teaching, English Literature, and Translation Studies programs at M.A. level in Iran. For the purpose of the present illustration only the reading comprehension data of the GE part of the UEE were used.

Cognitive Diagnostic Models

Cognitive diagnostic analysis promotes assessment *for* learning and the learning process as opposed to assessment *of* learning outcomes (Jang, 2008). Through providing detailed diagnostic feedback, it can inform teachers to modify instruction and learning in classrooms, if needed. CDM is an interdisciplinary approach to diagnostic assessment. It is at the interface between cognitive psychology and statistical analysis. It investigates the relationship between the psychological processes and strategies underlying performance on items of a given test and the responses provided to those items through sophisticated statistical analysis.

CDMs are latent class models (Haagenars & McCutcheon, 2002) that classify test takers into some latent classes according to similarity of their responses to test items. They are called *restricted* latent class models because the number of the latent classes is restricted by the number of attributes involved in answering items of a test. With K attributes underlying performance on a given test, the respondents will be classified into 2^K latent classes (the number 2 indicates that there are two possible outcomes for each attribute: mastery or nonmastery). In the case of the present study, for example, with five attributes required to perform successfully on the items of the test under study, test takers were classified into $2^5 = 32$ latent classes.

CDMs predict probability of an observable categorical response from unobservable (i.e., latent) categorical variables. These discrete latent variables have been variously termed as *skill*, *subskill*, *attribute*, *knowledge*, and *ability*. In the present paper, the terms attribute, skill, and subskill are used interchangeably to refer to the discrete latent predictor variables.

CDMs have been defined by Rupp and Templin (2008) as “probabilistic, confirmatory multidimensional latent variable models with a simple or complex loading structure” (P. 226). They are probabilistic models in that each CDM expresses a given respondent’s performance level in terms of the probability of mastery of each attribute separately, or the probability of each person belonging to each latent class (Lee and Sawaki, 2009). Like confirmatory factor analysis models, CDMs are also confirmatory in nature in the sense that latent variables in CDMs are defined a priori through a Q-matrix. A Q-matrix (Tatsuoka, 1985) is the loading structure of a CDM. It is a hypothesis about the required skills for getting each item right (Li, 2011). It is a matrix of as many rows as there items on the test and as many columns as there are attributes underlying performance on the test. CDMs are also multidimensional latent variable models because, unlike IRT models which assign to respondents a single score on a continuous scale, they assign respondents to multidimensional skill profiles by classifying them as masters versus non masters of each skill involved in the test. However, CDMs are notably different from multidimensional IRT models in that the latent variables in CDMs are discrete or categorical (e.g., masters/non-masters), whereas ability estimates (θ) in multidimensional IRT models are continuous. Finally, for the purpose of the CDMs, each item typically requires more than one attribute. This leads to a complex loading structure where each item is specified in relation to multiple attributes. This complex loading structure, in terms of multidimensional IRT, is known as *within-item multidimensionality* (e.g., McDonald, 1999).

CDM vs. IRT

Snow and Lohman (1989), by documenting the ways conventional educational psychometric measurement (EPM) models such as IRT are limited, tacitly pointed to the ways successful use of CDMs can overcome those limitations: They (a) explain item responses through a substantive psychological theory, (b) explicitly delineate the psychological processes that collectively underlie the construct measured by a test, (c) make realistic assumptions about the variables that affect performance on items of a test (as opposed to conventional models such as the three-parameter IRT model which makes a simplifying assumption that only three parameters affect item responses). More importantly, unlike conventional EPMs such as IRT, which are based on an investigator’s *expectations* of what cognitive processes test takers follow to solve problems in test taking situations, CDMs are based on empirical evidence of the actual processes and strategies followed in these situations.

All the EPM models aim to provide information about position of test takers along (a) latent variable(s) underlying performance in any assessment situation. Conventional IRT models locate test takers on a broadly defined single latent variable, whereas CDMs provide information about mastery status of test takers of a set of interrelated separable attributes. Mastery status is expressed either in terms of probabilities for each person having mastered each separate skill involved in answering successfully items of a test or in terms of a vector of 0/1s indicating nonmastery and mastery, respectively. In a test requiring four subskills, for example, a person who has mastered the first two attributes but not the last two, might be assigned the vector (1,1,0,0) or (.91, .86, .27..32), where 0s and probabilities below.5 indicate nonmastery, and 1s and probabilities above .5 indicate mastery.

Types of CDMs

Generally speaking, CDMs can be grouped into three families, as shown in Table 1:

Table 1. CDM Types

CDM Type	Examples	Author(s)
Compensatory	1) deterministic-input, noisy-or-gate model (DINO) 2) compensatory reparameterized unified model (C-RUM)	Templin & Henson (2006) Hartz (2002)
Non-compensatory	1) deterministic-input, noisy-and-gate model (DINA) 2) noncompensatory reparameterized unified model (NC-RUM)	Junker & Sijtsma (2001) DiBello et al. (1995); Hartz (2002)
General	1) general diagnostic model (GDM) 2) log-linear CDM (LCDM) 3) generalized DINA (G-DINA)	Von Davier (2005) Henson, Templin & Willse (2009) de la Torre (2011)

In compensatory models, mastery of one or some of the attributes required to get an item right can compensate for nonmastery of the other attributes. On the contrary, in noncompensatory models lack of mastery of one attribute cannot be completely compensated by other attributes in terms of item performance; that is all the attributes must function in conjunction with each other to produce the correct answer. General CDMs allow for both types of relationships within the same test. Many specific CDMs such as DINA, DINO, NC-RUM, C-RUM, and ACDM can be derived from the GDINA, for example. Thus GDINA allows a different model for each item on the same test. For one item, for example, the DINA model may be the best choice, for another the DINO, and still for the other the C-RUM.

The probability in a GDINA model that student *i* gets item *j* correct which requires two attributes α_1 and α_2 is defined as in Equation 1:

$$P(X_{ij} = 1|\alpha_1, \dots, \alpha_K) = \delta_{j0} + \delta_{j1}\alpha_1 + \delta_{j2}\alpha_2 + \delta_{j12}\alpha_1\alpha_2 \quad (1)$$

The parameter δ_{j0} is denoted as the item intercept which is the probability of a correct answer to an item when none of the required attributes for the item has been mastered. For two attributes, there are two main effects δ_{j1} and δ_{j2} and one interaction effect δ_{j12}.

For three required attributes α_1, α_2 and α_3, the probability is defined as in Equation 2:

$$P(X_{ij} = 1|\alpha_1, \dots, \alpha_K) = \delta_{j0} + \delta_{j1}\alpha_1 + \delta_{j2}\alpha_2 + \delta_{j3}\alpha_3 + \delta_{j12}\alpha_1\alpha_2 + \delta_{j13}\alpha_1\alpha_3 + \delta_{j23}\alpha_2\alpha_3 + \delta_{j123}\alpha_1\alpha_2\alpha_3 \quad (2)$$

For the general formulation of the probability in the GDINA model see de la Torre (2011).

Steps in conducting CDM

CDMs have been employed in two ways: (a) *retrofitting* (*post-hoc analysis*) of existing non-diagnostic tests to extract richer information and (b) designing a set of items or task from the beginning for diagnostic purposes. Many of the applications of the CDMs (including the present illustration) in educational measurement in general and language testing in particular are cases of *retrospective* specification (post-hoc analysis) of the knowledge and skills evaluated on existing non-diagnostic tests.

The following steps are involved in retrofitting CDMs to existing tests:

1) Specifying the skills and attributes required to perform successfully on any given test. This stage is the personification of the *construct representation* stage of construct validation, proposed by Embretson (1983).

   To define attributes involved in a test, various sources such as test specifications, content domain theories, analysis of item content, think-aloud protocol analysis of examinees' test taking process, and the results obtained by the relevant research in the literature can be sought (Embretson, 1991; Leighton & Gierl, 2007; Leighton, Gierl, & Hunka, 2004). In specifying the subskills underlying any given test, some considerations need to be taken into account. Models with large average number of attributes per item are more likely to be *unidentified*. The finer the *grain size* of a CDM, the richer the diagnostic information provided (Alderson, 2005), however, the more stress is placed on the capacity of statistical modeling. Hartz, Roussos, and Stout (2002) suggested there should be at least three items associated with each attribute for diagnostically reliable information. As Lee and Sawaki (2009b) argued, the more detailed the level of specification of a Q-matrix, the larger the number of required items to represent the universe of the attributes in a test. They further argued that "In addition, it is likely that the more fine-grained the attributes are, the more difficult it can become to maintain the consistency of diagnosis across occasions or test forms, potentially contributing to instability and unreliability of examinee classification"(p.184).

2) Analysis of the test items and delineating skill-by-item relationships in a Q-matrix. According to Lee and Sawaki (2009b) the diagnostic power of a CDM depends on the theoretical and empirical soundness of a Q-matrix.

3) Model specification. The relationships (e.g., conjunctive, compensatory, or general) among the postulated subskills should be specified. Selection of an appropriate CDM which is suitable for a particular assessment purpose is a prerequisite in cognitive diagnostic analysis. Rupp and Templin (2008) discuss the confirmatory nature of CDMs in a way that is rarely noticed. They argue that CDMs are confirmatory in that the appropriate CDM which reflects how attributes interact in the response process (i.e., how mastery of the postulated attributes affects response probabilities) should be specified a priori.

4) Estimating the profiles of skill mastery for individual examinees based on actual test performance data using the CDM.

The Q-matrix used in the present illustration was adopted from a study by Ravand (2015) wherein he specified the subskills and developed the Q-matrix underlying the reading section of the UEE. Since the test employed in the study by Ravand had not been developed for diagnostic purposes, he took the following steps to ensure, as much as possible, that the subskills identified were valid:

(a) The author invited two university instructors to brainstorm on the possible attributes measured by the test,

(b) Three other university instructors and three Master students were invited to independently specify the attributes measured by each item,

(c) The Q-matrix was empirically validated and revised. There are a few methods available which have been developed to identify Q-matrix misspecifications (e.g., methods developed by Barnes, 2010; Chiu, 2013; DeCarlo, 2012; de la Torre, 2008; Liu, Xu, & Ying, 2012; Templin & Henson, 2006). The above mentioned methods are limited in that they have been applied to specific CDMs, one or another. To apply these methods, one has to make a priori specification of the model: compensatory or noncompensatory. Some of these methods such as the ones developed by Barnes (2010) and Liu et.al. (2012) are further limited in that they derive Q-matrices solely based on test takers' responses without taking into account expert opinion. de la Torre and Chiu (2010) proposed a discrimination index that can be used with all the specific CDMs that are subsumed under the G-DINA model. Ravand (2015) employed the same method to identify misspecifications of (to validate) the Q-matrix adopted in the present study.

(d) The final Q-matrix was cross-validated with the other half of the sample[1]. According to the Q-matrix construction phase of the study, there were five attributes underlying performance on the reading comprehension section of the UEE: *reading for details*, *reading for inference*, *reading for main idea* (henceforth referred to as Detail, Inference, and Main Idea, respectively) *Syntax*, and *Vocabulary*. For a detailed process of Q-

[1] Ravand (2015) split the sample for his study into two: Half of the sample was used to identify and revise Q-matrix misspecifications in Stage C and the other half was used to cross-validate the Q-matrix thus obtained.

matrix development and revision refer to Ravand (2015).

As to the third stage (i.e., model specification) in CDM analysis, the authors chose the G-DINA to demonstrate.

Benefits of using R to conduct CDM

Most, if not all, of the software programs available to estimate CDMs handle only one of the CDMs shown in Table 1. MDLTM (von Davier, 2006), for example, conducts GDM, Arpeggio Suite (Bolt et. al., 2008) conducts the Fusion model (NC-RUM), and Mplus (Muthen & Muthen, 2013), which is a general purpose software, can conduct LCDM. Arpeggio and Mplus are commercial software but a restricted research license of GDM is available from the author free of charge. Another problem with some of the most commonly used software like Arpeggio and MPlus is that preparing their syntax is a tedious process and especially in the case of Mplus it involves minute specifications. For a four-attribute test, for example, several pages of syntax must be written for Mplus. To ease the pain of heavy syntax building, Templin and Hoffman (2013) have prepared a SAS macro that automatically generates the required syntax for Mplus to conduct LCDM but model estimation is carried out through Mplus. Furthermore, software such as Arpeggio and Mplus are relatively time inefficient in estimating CDM parameters. As the number of the attributes involved in a test increases, the time taken to estimate the model exponentially increases. Each run of the Arpeggio with its default number of Markov Chain Monte Carlo chains and four attributes, for example, would take about 28 minutes (for 1500 subjects and 1000 iterations) on a computer with 2GB of RAM and a Core i3 CPU. Depending on how many times a researcher revises the Q-matrix, she would spend hours estimating the model parameters with the software. As it was mentioned before, although syntax generation is carried out by the SAS macro, LCDM parameter estimation is carried out through Mplus, which would take several hours with four or five attributes. After all, depending on the nature and purpose of the study, a researcher may want to run more than one CDM. For example, he may want to compare a compensatory CDM with a noncompensatory one to ensure about the nature of the relationship among the attributes underlying a given test. To this end, she or he has to buy and learn how to work with more than one software programs, which would be a burden on the researcher both financially and technically.

The four most appealing features employing the R package CDM (Robitzsch, Kiefer, Cathrice George, & Uenlue, 2014), intended for cognitive diagnostic analysis, are: (a) It is very time efficient: Estimation of the parameters of anyone of the above mentioned models with, for example, five attributes would take less than a minute, (b) It has the capability to run most of the major CDMs such as DINA, DINO, NC-RUM, GDM, and G-DINA, (c) It is free, and (d) Anyone of the CDMs can be conducted with just a few lines of syntax.

Working with R

As with any other analysis in R, before conducting CDM analysis, the relevant package should be loaded. The CDM package is loaded by executing the following command:

```
library(CDM)
```

In order to conduct CDM, two data files are required: A file that embodies test takers' responses to the items of a given test and a file which includes the Q-matrix. A portion of the Q-matrix used for the illustration purpose in this study is displayed in Table 2. In a Q-matrix 1s indicate that *k*th attribute is required by *i*th item and 0s indicate that the attribute is not required by the item. For example, as Table 2 shows, Item 1 requires Inference and Vocabulary attributes whereas Item 5 requires only Vocabulary.

Table 2. Q-matrix

Item	Detail	Inference	Main idea	Syntax	Vocab
1	0	1	0	0	1
2	1	0	0	0	1
3	0	1	0	0	1
4	0	0	1	1	1
5	0	0	0	0	1

After the package has been loaded, the data should be imported into R. The most convenient way to import data into R, is to save the file in the format of comma separated values (`.csv`) or tab-delimited text (`.txt` or `.dat`).

The foreign package can be used to import data from different statistical packages (like SPSS or Stata) into R. To do so, execute the following command:

```
library(foreign)
```

Depending on the format of the data `read.delim` or `read.csv` functions can be used to assign the data to an object as follows:

```
data <- read.csv(“filename.extension”,
header=TRUE)
```

`header=TRUE` tells R to read the variable names form the first row of the data file. If the data file does not have the variable names, the argument `header=FALSE` should be used.

Now import the data file and Q-matrix into R. The data file and the Q-matrix had been saved under the names of “san.csv” and “qmat.csv”, respectively. Therefore, they can be imported by executing the following commands:

```
mydata <- read.csv("san.csv", header=TRUE)
[,c(2:21)]

qmat <- read.csv("qmat.csv", header=TRUE)
```

Since item data are located in Columns 2 to 21, the brackets at the very end of the first command select all the rows and Columns 2 to 21 (within the brackets, what comes before comma refers to rows and what comes after comma refers to columns)

As it was mentioned before, the CDM package is capable of conducting several CDMs such as DINA, DINO, NC-RUM, ACDM, GDM, and GDINA. For the purpose of the present paper, GDINA is illustrated.

The main function that estimates GDINA is

```
gdina(data, q.matrix)
```

For example, in our case with the previously created objects of 'mydata' and 'qmat' (we could have given other names to the objects created) the function becomes

```
model1 <- gdina(mydata, qmat)
```

Note we have created the object `model1` from the application of the `gdina` function to the two objects of `mydata` and `qmat`.

When the estimation finished, execute the following command to get GDINA item parameters:

```
model1$probitem
```

The dollar sign (`$`) in R codes means that the operation specified after `$` should be called within the object named before `$`. This command tells R to extract the coefficients from the object `model1`.

To save the output in a file, `write.csv` or `write.delim` functions can be used as follows:

```
write.csv(model1$probitem,
file="gdinparam.csv")
```

By executing this command we ask R to write the output of the `model1$probitem` function in a `csv` format file which we named it `gdinparam`. The result will be saved in an excel file. Part of the output is displayed in Table 3.

Table 3. G-DINA Parameters

Itemno	partype.attr	attributecomb	prob
1	V2-V5	A00	.11
1	V2-V5	A10	.18
1	V2-V5	A01	.42
1	V2-V5	A11	.60
2	V1-V5	A00	.10
2	V1-V5	A10	.19
2	V1-V5	A01	.46
2	V1-V5	A11	.53

Note. V1 to V5 are Detail, Inference, Main idea, Syntax, and Vocabulary, respectively. Itemno: item number; parttype.attr: Attributes required; attributecomb: attribute combinations; prob: probabilities

In this table the second column represents the attributes required by any item, the third column displays the attribute mastery patterns and the fourth column represents the probability of success on each item due to mastery of attributes required by the item. The number of parameters estimated for each item is a function of the number of attributes required by that item. Since G-DINA is a saturated CDM, all the main effects for the attributes and all their possible interactions are estimated.

As Table 3 shows, those who have not mastered any of the attributes required by Item 1 (indicated by the pattern A00) namely Inference (V2) and Vocabulary (V5), have about 11% chance of *guessing* and getting the item right. Chances of success on Item 1 for those who have mastered only Inference (indicated by the pattern A10), were 18% higher compared to those who have not mastered any of the attributes. Therefore, masters of Inference had .11 +.18 = .29 probability of *not slipping* (success) on the item. Mastery of Vocabulary (indicated by the pattern A01) increased success on the item more than mastery of the Inference, indicating that Attribute 2 discriminated more between its masters and non-masters. Therefore, masters of vocabulary had .11+.42=.53 chance of getting the item right. Interaction of (mastery of both) Attributes 1 and 2 added 60 % to the probability of success on the item. For masters of both attributes (indicated by the pattern A11) the probability of getting the item right was .11+.60=.71.

To obtain the attribute class probabilities, the following command should be executed:

```
model1$attribute.patt
```

Table 4. Class Probabilities

Latent class	Attribute profile	Class probability	Class expected frequency
1	00000	.149	3223.2
2	10000	.001	18.1
3	01000	.004	81.4
…	…	…	….
31	01111	.018	388.7
32	11111	.373	8073.9

Table 4 shows a portion of the output generated by executing the above command. In the present study test takers were classified into 2^5=32 latent classes. The second column of the table shows the possible attribute profiles for all the 32 latent classes. As the third column of Table 4 shows, the attribute profile of α_{32}=[11111] had the highest class probability of about .37. Approximately, 37% of the respondents (as shown in the last column, about 8073 respondents) in the present study were classified as belonging to this last latent class hence expected to have mastered all of the five attributes. Attribute profile of α_1=[00000] had the second highest class probability of about .15 indicating that approximately 15% (about 3223 respondents) of the test takers were expected to have mastered none of the attributes.

To obtain probabilities for each respondent belonging to any of the 32 latent attributes, execute the following command

```
model1$posterior
```

and you will obtain the output shown in Table 5. Table 5 has been transposed to fit the printed page.

Table 5. Class Probabilities for Respondents

	Response pattern		
	0000000 0000000 000000	1110010 0001000 100000	0100000 0010110 010000
Class 1	0.98	0	0
Class 2	0	0	0
Class 3	0	0	0
Class 4	0	0	0
Class 5	0	0	0
Class 6	0	0	0
Class 7	0	0	0
Class 8	0	0	0
Class 9	0.01	0	0.02
Class 10	0	0	0.02
Class 11	0	0.03	0.02
Class 12	0	0.01	0.06
Class 13	0	0	0
Class 14	0	0	0
Class 15	0	0.08	0.07
Class 16	0	0.06	0.47
Class 17	0	0	0
Class 18	0	0	0
Class 19	0	0	0
Class 20	0	0	0
Class 21	0	0	0
Class 22	0	0	0
Class 23	0	0	0
Class 24	0	0	0
Class 25	0	0.08	0.03
Class 26	0	0	0.02
Class 27	0	0.07	0
Class 28	0	0.07	0.04
Class 29	0	0	0
Class 30	0	0	0
Class 31	0	0.16	0
Class 32	0	0.43	0.23

Table 5 displays the probabilities that each person belonged to anyone of the 32 latent classes, for three respondents. In the table, values for each respondent with the given response pattern represent the posterior probability that he belonged to latent class *c* with the given attribute profile. For example, for Respondent 2, the chances were 43 % and 16% that she or he belonged to latent classes 32 and 31, respectively. Put another way, there is 43% chance that he has mastered all the five attributes and 16% chance of having mastered Attributes of Inference, Main idea, Syntax, and Vocabulary.

To obtain probabilities that each test taker has mastered any of the attributes involved in answering the items of the test, execute the following command:

```
model1$pattern
```

To save space, only a portion of the output is presented in Table 6. It shows the probability that each respondent with the given ID, response pattern, and attribute profile has mastered Attributes 1 to 5. For example, the probabilities that Respondent 6085 with the attribute profile of $\alpha_{25}=$ [10010] has mastered Attributes1 to 5 were .47, .84, .90, .78 and 1.00, respectively.

Model Fit

Like in any other statistical model, estimated parameters in CDMs are interpretable to the extent that the model fits the data. Fit of a model can be ascertained in two ways: checking fit of the model to the data (i.e., *absolute fit*) and comparing the model with other rival models (i.e., *relative fit*). The CDM package generates a range of absolute fit indices by comparing the observed and model-predicted response frequencies of item pairs (Maydeu-Olivares, 2013).

R generates absolute and relative fit indices by executing the following command:

```
IRT.modelfit( model1)
```

Table 6. Attribute Mastery Probabilities

Column1	**pattern**	**attribute profile**	**Prob-ability**	**attribute1**	**attribute2**	**attribute3**	**Attribute4**	**Attribute5**
1	00000000000000000000	00000	.98	.00	.00	.00	.01	.00
14238	11100100001000100000	01011	.43	.58	.91	.74	1.00	.82
6085	01000000010110010000	10010	.47	.84	.90	.78	1.00	.32

Difficulty of the attributes can also be calculated. Executing the following command will return the percentage of subjects who have mastered each attribute.

```
model1$skill.patt
```

Table 7. Attribute difficulty

Atttribute	Attribute.prob
Detail	.60
Inference	.50
Main idea	.54
Syntax	.72
Vocabulary	.64

As Table 7 shows, Syntax, mastered by about 73 % of the test takers, was the easiest attribute followed by Vocabulary, Detail, Main Idea, and Inference mastered by 64 %, 60 %, 54 % and 50 % of the test takers, respectively. Therefore, Syntax was the easiest and Inference was the most difficult attribute.

The model fit indices are presented in Tables 8a and b. Table 8a includes relative fit indices of information criteria AIC, BIC, AIC3, sample size adjusted AIC (AICc) and consistent AIC (CAIC). The model with the least information criteria is the most preferable. It turns out that the GDINA model fits the data best with respect to all criteria. Besides these measures of relative model fit, the `IRT.modelfit` function also provides a significance test of absolute model fit (maxX2; see Chen, de la Torre & Zhang, 2013). As Table 8a shows, the least value was obtained for the GDINA model (maxX2 = 20.26), however there was a significant model misfit (p = .001). The DINA model and the ACDM had a worse model fit.

Like in structural equation modeling, effect sizes of absolute model fit have been proposed (Robitzsch et al., 2014). The CDM package especially provides measures MADcor, SRMSR and 100*MADRESIDCOV (MADRCOV) which compare observed and predicted covariances (or correlations) of item pairs. The smaller an effect size, the better a model fits. From the results of MADRCOV (Table 8b), we conclude that the GDINA model (MADRCOV=.123) and the ACDM model (MADRCOV=.162) were clearly superior to the DINA model (MADRCOV=.431).

Table 8a. Fit Indices

Model	loglike	Deviance	Npars	Nobs	AIC	BIC	AIC3	AICc	CAIC	maxX2
gdin	-234905	469811	82	21642	469975	470629	470057	469976	470711	20.26
dina	-237461	474923	56	21642	475035	475482	475091	475035	475538	270.29
acdm	-235249	470499	68	21642	470635	471178	470703	470635	471246	170.97

Table 8b. Fit Indices

Model	p_maxX2	MADcor	SRMSR	100* MADRESIDCOV (MADRCOV)	MADQ3	MADaQ3
gdin	0.001	0.006	0.007	0.123	0.025	0.020
dina	0	0.020	0.028	0.431	0.022	0.021
acdm	0	0.008	0.010	0.162	0.027	0.022

The `IRT.modelfit` function also performs a likelihood ratio test for model comparisons which are valid when the models under study are nested. The output is displayed in Table 9. It is evident, that the GDINA model fitted the data significantly better than the DINA model (Chi2(df=26)=5112.10, p<.001) and was also superior to the ACDM model (Chi2(df=14)=688.03, p<.001).

Table 9. Model Comparison

$LRtest	Model1	Model2	Chi2	df	p
1	dina	gdin	5112.10	26	0
2	acdm	gdin	688.03	14	0
3	dina	acdm	4424.07	12	0

Differential item functioning

Another capability of the CDM package is that it can also perform differential item functioning (DIF) in the context of CDM. According to Hou, de la Torre, and Nandakumar (2014, p.99) "In the context of CDMs, DIF is an effect where the probabilities of correctly answering an item are different for examinees with the same attribute mastery profile but who are from different observed groups". Unlike traditional DIF detection procedures which use the total score as the matching criterion, the CDM DIF detection procedure proposed by Hou et al. uses attribute mastery profile score as the matching criterion. The procedure has the following advantages: (a) It can investigate both uniform and nonuniform DIF, (b) Item calibrations are done separately for the reference and focal groups through the Wald test thus contamination due to DIF items is avoided and the need for purifications is obviated. In the CDM DIF, uniform DIF exists when probability of answering an item is the same for test takers of one group across all the attribute profiles. If this probability changes for test takers of the same group across the attribute profiles (i.e., higher on some attribute profiles but lower on the others) there is an indication of nonuniform DIF. To conduct DIF in the CDM package, one needs to fit a multiple group G-DINA first. To introduce gender as the grouping variable, the following command should be executed:

```
multigdin<-gdina( mydata, qmat , group =
data$gender )
```

Finally, the following function can be employed to conduct CDM DIF:

```
difres <- gdina.dif(multigdin)
```

The output can be recalled by the following code:

```
summary(difres)
```

The third column of Table 10 shows that the difficulties of Items 55 and 57 were significantly different for males and females (*p* <.05). The last column (i.e., UA) shows the effect size for DIF. Jodoin and Gierl (2001) suggest as a rule of thumb values of .059 to distinguish negligible from moderate DIF and .088 to distinguish moderate from large DIF. As Table 10 shows, the effect size for both items flagged for DIF are blow .059. Therefore, it can be concluded that the very high sample size of the present study rendered the small differences in the difficulty of the respective items between males and females statistically significant.

Table 10. DIF estimates

Item	χ^2	*df*	*P*	UA
41	2.08	4	.72	.035
42	1.16	4	.88	.020
43	1.44	4	.84	.011
44	1.07	4	.90	.013
45	0.07	2	.97	.003
46	0.20	4	.99	.008
47	0.39	4	.98	.016
48	1.52	4	.82	.016
49	0.65	4	.96	.014
50	4.48	2	.11	.028
51	4.37	2	.11	.006
52	3.22	8	.92	.033
53	1.15	4	.89	.025
54	0.16	2	.92	.008
55	6.64	2	.04	.019
56	1.51	2	.47	.013
57	8.12	2	.02	.026
58	2.04	2	.36	.016
59	2.49	4	.65	.039
60	0.04	2	.98	.003

Discussion

In this paper we first reviewed CDMs and showed how CDM package in R can be conveniently used to conduct cognitive diagnostic analysis. We also briefly introduced the R environment. We guided the reader through the steps required to do CDM. We also provided an accessible easy-to-understand interpretation of the output of G-DINA.

Applications of CDMs have mainly focused on classifying test takers into multidimensional skill spaces, thereby providing detailed diagnostic information of strengths and weaknesses of test takers (e.g., Buck & Tatsuoka, 1998; Jang 2009a; Kasai, 1997; Kim, 2011; A. Kim, 2014; Li, 2011; Li & Suen, 2013; Ravand, Barati, & Widhiarso, 2012; Sawaki, Kim, & Gentile, 2009; von Davier, 2005). The information provided by CDMs can have theoretical and practical implications as well. Theoretically, as de la Torre and Lee (2013) noted, specific CDMs for each item can indicate how attributes underlying a test can combine (e.g., in a compensatory or conjunctive way) to produce correct responses to items of the test. Practically, they can be used to explore what features make the items conjunctive or compensatory. This function of the CDMs is of interest especially when items requiring the same attributes have different inter-skill relationships.

CDMs can also be employed to demonstrate, in Chronbach and Meehl's (1955) word, *strong* form of construct validity. As Rupp and Templin (2008) discussed, CDMs are confirmatory in two ways: First, according to a substantive theory of a construct, the knowledge and processes which test takers require to perform successfully on the items of a test are described in a Q-matrix. Using an analogy from confirmatory factor analysis, we can say that a Q-matrix is the loading structure of a CDM wherein item-by-skill relationships are hypothesized. Then the theory-driven Q-matrix is validated against real data. According to Rupp and Templin, CDMs are also confirmatory in that how attributes interact in the response process should be specified a priori, that is whether attributes combine in a compensatory or conjunctive relationship to produce the correct answer should be specified in advance. The process of model selection is informed by the domain theories or the extant literature. Therefore, "as with most procedures for validating theories in scientific investigations, model selection is conducted by comparing the theory-based predictions and actual observations" (de la Torre and Lee, 2013, p.356). Thus when a compensatory or a conjunctive model is selected to explain the relationships between the attributes and item response probabilities, if characteristics of the data can be reproduced by the model, it is said that the model fits the data hence the postulated relationships are confirmed. From both *skill specification* and *model selection* perspectives, CDMs involve theory testing, which is what Chronbach and Meehl's (1955) strong program of validity entails.

References

Bolt, D., Chen, H., DiBello, L., Hartz, S., Henson, R. A., & Templin, J. L. (2008). *The Arpeggio Suite: Software for Cognitive Skills Diagnostic Assessment* {Computer software and manual}. St. Paul, MN: Assessment Systems .

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.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123-140.

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52,281–302.

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

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115-130. doi: 10.3102/1076998607309474

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. (2010). A general method of empirical Q-matrix validation using the G-DINA model discrimination index. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO

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

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.E. (1983). Construct validity: construct representation vs. nomothetic span. *Psychological Bulletin*, 93, 179–197.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495-515.

Haagenars, J., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.

Hartz, S. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practice.* Unpublished doctoral thesis, University Illinois at Urbana-Champain .

Hartz, S., Roussos, L., & Stout, W. (2002). Skills diagnosis: Theory and practice. *User Manual for Arpeggio software. ETS.*

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191-210. Doi/10,1007: s11336-008-9089-5

Hou, L., la Torre, J. d., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98-125.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. *Towards adaptive CALL: Natural language processing for diagnostic language assessment*, 117-131.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349.

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.

Kim, A. Y. A. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, 0265532214558457.

Kim, Y. H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified model. *Language Testing*, 0265532211400860.

Lee, Y.-W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172-189. *doi: 10.1080/15434300902985108*

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

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

Leighton, J. P., & Gierl, M. J. (2007). 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.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach. *Journal of Educational Measurement*, *41*(3), 205-237.

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

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). *Measurement: Interdisciplinary Research and Perspectives, 11*, 71-137.

Muthén, B. O., & Muthén, L. K. (2013). *Mplus Version 7.11: User's Guide*. Los Angeles, CA: Muthen & Muthen.

Ravand, H. (in press). Assessing testlet effect, impact, differential testlet and item functioning using cross-classified multilevel measurement modeling. *SAGE Open.*

Ravand, H. (2015). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. Manuscript submitted for publication.

Ravand, H., Barati, H., & Widhiarso, W. (2012). Exploring Diagnostic Capacity of a High Stakes Reading Comprehension Test: A Pedagogical Demonstration. *Iranian Journal of Language Testing*, 3(1).

Robitzsch, A., Kiefer, T., George, A., C., & Uenlue, A. (2014). *CDM: Cognitive Diagnosis Modeling.* R Package Version 4.1. http://CRAN.R-project.org/package=CDM

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

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: American Council on Education/Macmillan.

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*(3), 287-305. doi: 10.1037/1082-989x.11.3.287

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37-50. doi: 10.1111/emip.12010

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

von Davier, M. (2006). *Multidimensional latent trait modelling* (MDLTM) [Software program]. Princeton, NJ: Educational Testing Service.

Citation:

Ravand, Hamdollah & Robitzsch, Alexander (2015). Cognitive Diagnostic Modeling Using **R** *Practical Assessment, Research & Evaluation*, 20(11). Available online: http://pareonline.net/getvn.asp?v=20&n=11

Corresponding Author:

Hamdollah Ravand
Vali-e-Asr University of Rafsanjan
Iran

Email: ravand [at] vru.ac.ir