Models for Cognitive Diagnostic Modeling

Phonraphee Thummaphan[1],
Min Li[2],
Jimmy de la Torre[3]

Abstract

Psychometric techniques have been developed to measure cognitive constructs and provide practical implications. Cognitive diagnostic modeling (CDM) is a measurement method that is useful for diagnosing the cognitive mastery of examinees. CDM involves various models and steps. We start by providing an overview of CDM. Then we describe four psychometric models for cognitive diagnosis that are most commonly encountered in the literature by discussing their assumptions, strengths, and weaknesses, and the extent to which they relate to each other. Also, we discuss the different steps involved in CDM (i.e., attribute definition, different types of validation, model fitting, and different approaches to model selection). Finally, we propose an integrative framework with which one may carry out CDM.

Keywords: psychometric techniques, cognitive diagnostic modeling

[1] University of Washington, Seattle, USA
[2] University of Washington, Seattle, USA
[3] Rutgers, The State University of New Jersey, USA

Overview of CDM

Definition of CDM

CDM which typically in values a cognitive diagnosis model (also known as diagnostic classification model), is a psychometric modeling approach that can be used to measure students' cognitive skills or knowledge required to answer items correctly. Cognitive Diagnostic Models are "latent variable models developed primarily for assessing student mastery and nonmastery on a set of finer-grained skills" (de la Torre, 2011). The goal of CDM is to measure student mastery on the required skills, and to evaluate the diagnostic power of item measuring such skills (Hartz, 2002). The main advantage of CDM is that they provide a detailed profile on student understanding by specifying strengths and weaknesses of individual learners rather than an overall score. Thus, assessment results can be used for diagnosis and improvement of student learning (Pellegrino, 2013).

Various CDM approaches have been developed, including the deterministic inputs, noisy 'and' gate and noisy-input, deterministic 'and' gate (DINA and NIDA; e.g., Junker & Sijtsma, 2001), multiple-choice DINA (MC-DINA; e.g. de la Torre, 2009a), multi-strategy DINA (MS-DINA; De la Torre & Douglas, 2008), deterministic inputs, noisy 'or' gate and noisy-input, deterministic 'or' gate (DINO and NIDO; e.g., Templin & Henson, 2006), reparametrized unified model (RUM/fusion; e.g., DiBello, Stout, & Roussos, 1995; Hartz, 2002), noncompensatory reparameterized unified model (NC-RUM; e.g., DiBello et al., 1995; Hartz, 2002), random effects reparameterized unified model (RE-RUM; e.g. Templin & Henson, 2005), multiple classification latent class model (MCLCM; e.g. Maris, 1999), general diagnostic model (GDM; e.g., von Davier, 2005; Xu & von Davier, 2006), log-linear cognitive diagnostic model (LCDM; e.g von Davier, 2005, 2014), nominal response LCDM (NR-LCDM; e.g. Templin et al., 2008), and generalized DINA (G-DINA; e.g. de la Torre, 2011).

Each CDM model has different characteristics. Rupp and Templin (2008) classified CDMs by jointly considering three characteristics: (1) measurement scales of observed response variables (dichotomous vs. polytomous), (2) measurement scales of latent predictor variables (dichotomous vs. polytomous), and (3) model type based on the combination rule of latent predictor variables (compensatory vs. noncompensatory). Gu (2011) made the classification more complete by adding the Nominal Response models to the observed response category. Characteristics of each model (Table 1) were determined by summarizing Rupp and Templin (2008) and Gu (2011), as well as by adding conjunctivity information.

Table 1 Characteristic of CDMs

Model	Observed Response Variables			Latent Predictor Variables		Combination Rule		Conjunctivity	
	dichotomous	polytomous	Nominal	dichotomous	polytomous	compensatory	non-compensatory	conjunctive	disjunctive
DINA	✓			✓			✓	✓	
MS-DINA	✓			✓			✓	✓	
NIDA	✓			✓			✓	✓	
MCLCM	✓	✓		✓	✓	✓	✓	✓	✓
NC-RUM	✓	✓		✓	✓	✓	✓	✓	

Model	Observed Response Variables			Latent Predictor Variables		Combination Rule		Conjunctivity	
	dichotomous	polytomous	Nominal	dichotomous	polytomous	compensatory	non-compensatory	conjunctive	disjunctive
rRUM	✓			✓			✓	✓	
DINO	✓			✓		✓			✓
NIDO	✓			✓		✓			✓
C-RUM	✓	✓		✓	✓	✓		✓	
GDM	✓	✓		✓	✓	✓		✓	
LCDM	✓	✓		✓	✓	✓		✓	
MC-DINA			✓	✓			✓	✓	
NR-LCDM			✓	✓		✓		✓	
G-DINA	✓			✓			✓		✓

As shown in Table 1, in addition to the types of observed response and latent predictor variables, the combination rule of latent predictor variables is used to classify the models. Rupp and Templin (2008) note that "the difference between compensatory and noncompensatory models reflects how the latent predictor variables are combined across the different skills to produce the observed responses." That is, compensatory models assume that a low value on or lack of one skill can be compensated for by a high value on another skill. In contrast, a noncompensatory model assumes that a low value on or lack of one skill cannot be compensated for by a high value on another skill.

One more assumption related to CDM is conjunctivity, or the condensation rule, which makes assumptions about how a correct response occurs. Conjunctive models assume that correct responses occur when the examinee masters all "required" attributes, while disjunctive models assume that mastery of one or more "required" attributes is sufficient to produce the correct response (Broaddus & Shaftel, 2012; Junker, 1999; Rupp, Templin, & Henson, 2008). However, Junker (1999) mentions that it remains unclear how the attributes combine to produce a response. Moreover, conjunctivity can be related to the combination rule as well. For example, disjunctive models can be viewed as compensatory because one attribute can compensate for others and produce the correct response, particularly for the completely compensatory case (Junker, 1999; de la Torre & Douglas, 2004).

It is important to note that the attribute hierarchy method (AHM; e.g., Leighton, Gierl, & Hunka, 2004) and rule-space methodology (RSM; e.g., Tatsuoka, 1983, 1995) which are usually found in the literature, are not really CDMs, but a framework. To illustrate, while models involve a direct statistical link between observed response variables and latent predictor variables, RSM and AHM (an extension of the RSM) do not. Thus, they "are essentially classification algorithms and not unified statistical models that are completely embedded within a fully probabilistic framework" (Rupp & Templin,

2008). Moreover, Bayesian inference networks (BINs; e.g., Yan, Mislevy, & Almond, 1993), is not a single model, as are the others, but is instead "a general modeling framework for representing different kinds of latent variable models" (Rupp & Templin, 2008).

Commonly Used Cognitive Diagnostic Models

A number of commonly used CDMs are related to each other (de la Torre & Douglas, 2004; de la Torre & Chiu, 2010; de la Torre, 2011; Rojas, de la Torre, & Olea, 2012). In this paper, we present four such models: DINA, NIDA, DINO, and G-DINA, with DINA as the focus of the paper and the model used as a point of comparison for the other models.

The DINA Model

The DINA Model (deterministic inputs, noisy ''and'' gate; e.g., Haertel, 1989; Junker & Sijtsma, 2001; de la Torre & Doulas, 2004; de la Torre, 2009b) is consided one of the most parsimonious and interpretable CDMs, as it requires only two parameters for each item regardless of the number of attributes involved (de la Torre, 2009; de la Torre, 2011; Lee, Park, & Taylan, 2011). The DINA model is a noncompensatory and conjunctive model. Taking the deterministic perspective, the DINA assumes that examinees who only have all the required attributes can answer items correctly. It also assumes that an examinee cannot compensate with any other attribute that he/she has, regardless of its high magnitude (Rupp et al., 2008). Thus, the DINA model classifies examinees into two classes on each item: (1) item masters, comprising those who have all required attributes to answer the item correctly, and (2) item nonmasters, comprising those who are otherwise (Rupp et al., 2008). However, if one takes the probabilistic view, examinees can still answer items correctly or incorrectly, based on what Rupp et al. (2008) call "lucky guesses" or "careless errors". Original terms are "guess" and "slip" (Junker & Sijtsma, 2001).

The DINA formula model includes three important components: the latent variable, slipping parameter, and guessing parameter (Rupp et al., 2008), as shown in the probability of a correct response of Examinee i who has the skills vector or latent class α_i for Item j as follows:

$$P_j(\alpha_i) = P\left(X_{ij} = 1 \,\middle|\, \alpha_i\right) = {g_j}^{1-\eta_{ij}}(1 - s_{ij})^{\eta_{ij}}$$

where $P_j(\alpha_i)$ is the probability of a correct response for item j in skills vector α_i, $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ is the binary indicator of examinee mastery status of all required attributes, X_{ij} is the observed response for Item j, α_i is the skills vector …, $(1 - s_{ij})$ is the probability of not slipping, and g_j is the probability of guessing.

It is important to understand the term deterministic inputs, noisy ''and'' gate, which describes the characteristics of this model. In this formula, η_{ij} is the latent variable that is the deterministic input of the DINA model used to indicate or classify examinee mastery for the item. The and-gate is the conjunctive process, indicating whether or not all required attributes are present, which creates the variable η_{ij}. The and-gate "functions like an output summary that takes the individual elements and condenses them much as a gate in a park forces people to enter and exit in a particular way" (Rupp et al., 2008). Noisy refers to slip and guessing parameters – slip indicates answering the item incorrectly due to a careless error despite the examinee having all required attributes, whereas guessing refers to answering the item correctly by guessing even though the examinee lacks the required attributes (Rupp, et al., 2008; de la Torre, 2009b).

Therefore, the DINA model provides one slip and one guessing parameter for each item, regardless of the number of attributes involved. That is, the total number of parameters estimated in the DINA model is determined by the number of items instead of the number of attributes (Rupp, et al., 2008). This makes the DINA model parsimonious, but it requires that examinees must master all the measured attributes in order to answer the items correctly. Again, this model does not discriminate

between those who lack different numbers of required attributes, thus making it very restrictive (de la Torre, 2013).

NIDA, DINO, and G-DINA

Table 2 compares NIDA, DINO, and G-DINA against DINA, demonstrating the various strengths and weaknesses of each model. Briefly, while being parsimonious, DINA is very restrictive. The G-DINA model is the more flexible since it permits examinees with different attribute patterns to have different probabilities to answer the item correctly. However, it involves more parameters, and estimation takes a longer time, and also requires a larger sample size (de la Torre & Lee, 2013). NIDA is more flexible than DINA; however, parameters are still constrained to be equal across items. DINO and DINA are equally restrictive, but in different ways. In the DINO model which is compensatory, parameters are still constrained to be equal across attributes, whereas in DINA, all attributes are required for answering items correctly.

Table 2 Comparison of Commonly Used Models

Model	Number of Parameter Estimates	Level of Parameter Estimation	Strengths	Weaknesses
DINA	2	Item	- parsimonious, interpretable - small sample size	- Restrictive: all attributes require the correct answer; cannot differentiate examinees who lack different attributes
NIDA	2xJ	Attribute	- parsimonious, interpretable - small sample size - can distinguish examinees who lack different attributes	- parameters are constrained across items; may not be flexible enough for some analyses.
DINO	2	Item	- parsimonious, interpretable - small sample size - allows compensation of attributes	- parameters are constrained across attributes; may not be flexible enough for some analyses
G-DINA	Can be more than 2[a]	Item	-general, more flexible	- requires higher number of parameters - requires larger sample size - takes time to converge

Note: a. The number of attributes depends on the number of attributes required for each item.

Steps Involved in CDM

Overall, CDM applies the latent class model approach by classifying examinees according to the attribute patterns – the combinations of attributes possessed or not possessed. Based on students' responses on items, CDM iteratively fits the model with a fixed number of classes, given the number of attributes set prior to the analysis, and evaluates the model fit.

CDM is a part of CDA1 (Cognitive Diagnostic Assessment), but it focuses on the psychometric modeling aspect. The important steps involved and presented in this section are about attribute definition, validation, model fitting, and model selection. In any case, to provide full implementation of CDA, researchers would need to determine the cognitive model, develop test items, and administer the test to collect real data.

Attribute Definition

One of the most important steps for CDMs is attribute definition, which is the process of determining the specification and definition of the skill(s) of interest. Roussos et al. (2010) explained

that "in general, this component of the implementation process considers how many and what kinds of skills or attributes are involved, at what level(s) of difficulty, and in what form of interaction." So this process requires a clear assessment purpose and contemporary cognitive or learning theory that informs the knowledge structures and processes that a learner would use in answering/successfully completing the questions/tasks.

The attribute definition, which is the document with the definition details of all attributes, is developed based on the cognitive model or relevant framework of diagnostic interest. For example, Lee et al. (2011) developed the attributes of TIMSS mathematics knowledge and skills from the 2007 TIMSS Mathematics Framework (Mullis et al., 2005). For each attribute, a clear description of knowledge and skills should be provided. The elements of the Q-matrix are usually determined based on experts' judgments.

Below is an example of the attribute definition developed by Thummaphan, Dong, and Li (2015). This definition is used to analyze mathematics word problems in Grade 4 level based on the US Common Core Standards. The definition includes the label and descitpion of the underlying understanding as well as a tip for coding word problems used in the study.

C1 (Whole Numbers)

A student who has mastered this attribute should be able to understand and apply basic concepts and operations in whole numbers:

Including addition, subtraction, multiplication, division, exponentiation, sign, absolute value, place value, rounding of 2- or 3- digit integer values, number patterns, comparison of magnitude of numbers, greatest common divisor, least common multiple, etc.

The Q-matrix

After developing the attribute definition, the next step is producing the Q-matrix, an item-by-attribute binary matrix. Providing the correct Q-matrix specification is not an easy task as it must take into account the combination and structure of skills (De Carlo, 2011). In general, the rows are the items in the test and columns are the attributes required to answer items correctly. The matrix must cover all items. Each item will be analyzed and coded based on the attribute definition to determine which attributes are required to answer each item correctly. The numbers indicating whether an attribute is required or not to answer an item correctly will appear in the Q-matrix as 1 or 0, respectively. The Q-matrix elements are usually determined based on experts' judgments and verified by researchers to determine inter-rates reliability. Therefore, the coders must be trained to fully understand the attribute definition. Low agreement requires reexamination of the attribute definition and reconsidering the coders' understanding of the attribute definition.

Methods of Estimation

Roussos et al. (2010) mention that marginal maximum likelihood estimation (MMLE) is the preferred estimation method. This method estimates the item and person ability parameters from a likelihood function by using either the expectation maximization (EM) or the Markov chain Monte Carlo (MCMC). The EM algorithm is a common method for estimating latent variable models and latent class models, and takes less time compared to MCMC (DiBello, Roussos, & Stout, 2007; Rupp, et al., 2010; von Davier & Yamamoto, 2004). However, if the number of latent classes and the set of constraints increases, EM estimation is computationally very intensive and complex (DiBello et al., 2007; Rupp, et al., 2010). Therefore, the EM algorithm has been used to estimate simpler models such as the DINA (de la Torre, 2009b; Liu, Douglas, & Henson, 2009). The details of each method follow.

EM focuses on the iteration procedure to arrive at the maximum likelihood parameter values. It has two important steps: expectation and maximization (Von Davier & Yamamoto, 2004; Rupp, et al., 2010). In the expectation step, sufficient statistics for the unobserved latent variable level data are accumulated by using Bayes with the expected values of the unknown attribute profiles. In the maximization step, the item parameter and structural parameter values can be estimated using the

posterior probabilities with the full data from the response and latent variables. EM then repeatedly updates this estimation until the difference in either item parameter estimates or the likelihood functions across iterations is small which is the algorithm convergence (Rupp, et al., 2010).

Unlike EM, MCMC applies the Bayesian estimation method for estimating the posterior distributions for all parameters by sampling these distributions instead of computing them (Rupp, et al., 2010). First, the Markov chains, a particular statistical sequence of random simulated dependent values of all the parameters, must be generated. Each step involves producing a set of generated values for each parameter from a specific distribution. Estimating the posterior distribution involves continuous running of large numbers on long chains (Burn-in). Then, the MCMC can be accomplished by using these posterior distributions to estimate the parameter values and standard errors. As mentioned by Roussos et al. (2010), the number of chains, total length of each chain, and the amount for the burn-in are critical decisions that researchers ought to make. If those are properly chosen, the model will converge well.

The examinee ability vector, α, must also be estimated, either by maximum likelihood or by Bayesian techniques (Roussos et al., 2010). Roussos et al. point out that the estimation implementation can have constraints that are consequently relevant to model restrictions and related to software code and programming expertise.

Model Fit Statistics

Model fit statistics vary according to the model estimation methods used. The absolute fit statistics determine how well the model fits the actual data, whearas the relative fit statistics compare the fits between competing models. The absolute fit statistics include the residuals between the observed and predicted proportion of correct individual items, between the observed and predicted Fisher-transformed correlation of item pairs (referred to as transformed correlation), and between the observed and predicted log-odds ratios of item pairs. Computation of both the statistics and the standard errors for each statistics is required to perform the test, which determines whether the residuals differ significantly from zero (Chen, de la Torre, & Zhang, 2013).

Similarly, the posterior predictive model (PPM) can be used to compare model-predicted data with observed data statistics when the Bayesian approach is used to estimate model parameters (Roussos et al., 2010). The mean absolute difference (MAD) between predicted and observed statistics should be small to confirm that the model fits well. A number of studies have successfully used PPM (e.g., Jang, 2008; Templin & Henson, 2006).

Relative fit statistics which include – 2 log-likelihood (–2LL), akaike information criterion (AIC), and bayesian information criterion (BIC), are computed as a function of the maximum likelihood (ML) (Chen et al., 2013), which is similar to methods using log-likelihood fit statistics, which compare the fits of competing models (Roussos et al., 2010). Von Davier (2005) used these statistics by comparing the fit of the compensatory MCLCM with that of the unidimensional two-parameter logistic model in the analysis of TOEFL data.

The Wald Test is also used in CDM to determine the item level fit statistics for the purpose of item-by-item model comparison. Specifically, it is used to statistically compare the fit between the saturated (G-DINA) and reduced (i.e., DINA, DINO, Additive CDM) models in the G-DINA model context in terms of the Type I error and power of the test (see de la Torre, 2011; de la Torre & Lee, 2013, for details). Then, the Wald test allows researchers to determine the best fit model for each item, rather than the best fit model for the whole test.

Validity

Validity evidence is a very important aspect of CDMs. Roussos et al. (2010) stated that validity research of the diagnostic instrument showing the statistical evidence has not progressed well. Two types of validity need to be considered: internal and external. Internal validity is established using data from the test itself to evaluate the validity of mastery classification, whereas external validity is established using external data (Roussos et al., 2010).

One key step for establishing internal validity is specification testing of the Q-matrix. Several studies investigated this issue (Rupp & Templin, 2008; De Carlo, 2011; de la Torre, & Zhang, 2013). Rupp and Templin simulated data with different misspecifications such as underfitting the Q-matrix (i.e., specifying 0s where there should be 1s), overfitting the Q-matrix (i.e., specifying 1s where there should be 0s), and providing a balanced misfit for the Q-matrix (i.e., exchanging 0s and 1s while controlling for the overall number of changes), to examine the effects of Q-matrix misspecifications on the slip and guessing parameter estimates and misclassification rates. They found that misspecification of a particular item overestimated the parameters of that particular items.

De Carlo (2011) used the posterior mode estimation (PME) with the DINA model in the fraction subtraction data. The results revealed some classification problems such as examinees who have zero total scores being classified as mastering most skills. He concluded that the Q-matrix misspecification can heavily affect the classification accuracy.

In addition to considering Q-matrix misspecification, IMstats and EMstats (Hartz and Roussos, 2008) can be computed for items and examinees, and used for classifying internal validity for the RUM model. To illustrate the process for IMstats, examinees would be classified into two groups: item master, indicating the examinee possess all required skills for the item, and item nonmaster if otherwise. Next, for each item, IMstats compared the average observed scores between item master and nonmaster groups. If the attribute definition and Q-matrix coding are right, the two scores should differ greatly, suggesting the model is valid. But if they are close, researchers should reconsider the attribute definition and Q-matrix coding. For EMstats, a similar procedure is followed, but for examinees instead of items.

Examining differential item functioning (DIF) is another way to ensure the validity of the results. In CDM, “an item exhibits DIF in the context of CDMs if the probabilities of success on the item are different for examinees who have the same attribute mastery profile but are from different groups” (Hou, de la Torre, & Nandakumar, 2014). Hou et al. used the Wald test to detect both uniform and nonuniform DIF in the DINA model and found that Wald test performance was comparable to or outperformed the Mantel-Haenszel (MH) and SIBTEST.

In addition, Ayers, Rabe-Hesketh, and Nugent (2013) performed MH tests using the predicted skill set profiles as matching criterion from the logistic and latent regression and standard DINA model. They found some DIF items with different models. Thus, allowing the differences of skill mastery probabilities between groups can provide DIF detection results as well.

With respect to external validity, Roussos et al. (2010) mentioned that very little research has been done with the IRT-based latent class models. But comparing pretest and posttest scores of diagnostic instruments with well-designed remediation can be shown as evidence of external validity. If the post-test score is higher than the pretest score, it means that the diagnostic assessment is valid (Jang, 2008; Roussos et al., 2010).

Another method for determining external validity is to use comparable test data with the same model to compare the results. If the classification results (i.e., percentage of the classification) are consistent between two assessment tests, the model is valid. For example, Templin and Henson (2006) compared the classification results of the DINO model using MCMC between two psychological tests: South Oak Gambling Screen (SOGS; developed by Lesieur & Blume, 1987) and Gambling Research Instrument (GRI; developed by Feasel, Henson, & Jones, 2004). After classifying the pathological gamblers, the results found that 89.2% of the classifications resulting from those two tests was consistent, indicating the model is valid. In addition, Cohen’s of 0.69 ($p < .0001$) showed substantial agreement beyond chance (Roussos et al., 2010).

Related to these steps, an integrative framework by which one may carry out CDM from start to finish, including examples of both general and specific models, is presented in the next section.

An Integrative Framework

In this section, we propose an integrative framework for implementing CDM in practice. In this framework, we specifically highlight several theoretical considerations.

CDM as a Theory-Driven Model

CDM is a substantively-grounded model by nature (DiBello & Stout, 2003). Hence, the theory-driven approach should be applied for CDM to gain valid and interpretable formative value of the assessment (Kane, 2006). The choice of cognitive model should be based on the assessment purpose. In terms of the assessment purpose, one can decide whether one wants to (1) choose existing test data that matches the cognitive model of interest, or (2) refine the test items to make them more sensitive to the chosen cognitive model (Roussos et al., 2010). However, it is important for the CDM that the items should be highly sensitive to the cognitive construct so that the linkage between the cognitive construct and the CDM is well-matched. Choosing from available items may be limited in terms of matching the construct of interest. Therefore, designing and developing good test items is the most appropriate approach. In this case, Henson and Douglas (2005) showed that using a heuristic, which is a simple algorithm used for item construction, in the item bank based on the CDM Information Index (CDI) results in higher classification rates than randomly constructed tests in both DINA and RUM models.

After deciding the assessment purpose, the attribute space can be defined and guided by the cognitive model. Cognitive models are "theoretical maps of how people learn and organize content knowledge" (Broaddus & Shaftel, 2012). There are several issues related to attribute definition: model and inferences, availability of the cognitive model, and numbers of attributes.

First, it is important to know that each cognitive model has its own capacity to make inferences about examinees' cognitive strengths and weaknesses. Leighton and Gierl (2007) examined three cognitive models in educational measurement: Test specifications/large-scale assessment, domain mastery/curriculum-based assessment, and task performance/cognitive diagnostic. The researchers looked at each model's capacity to (a) inform test item development for measuring cognitive processes of interest, and (b) support the construct validity of inferences made about students' cognitive processing. Of the three models, the task performance model was found to be most appropriate for the CDM because its goal is to assess cognitive strengths and weaknesses on the cognitive domain with the capacity to provide high psychological evidence. However, it can measure low range of skills. In contrast, although the cognitive model of domain mastery can cover a large range of content, it mostly focuses on behavior strengths and weaknesses, whereas the cognitive model of test specifications focuses only on ranking behavior mastery.

Secondly, attribute definition requires availability of the cognitive model. One challenge is the limitations of existing cognitive and learning theories, especially the cognitive models based on information processing theory, a topic of current interest in some disciplines such as problem solving in math and science. However, there are ways to develop the cognitive model or process and attribute validation, including verbal reports and protocol studies, eye-tracking research, and expert panels (de la Torre, 2014; Rupp, et al., 2010). Using evidence from these types of studies, the cognitive model can be used for the development of attribute lists and items. Such work must be well-established in order to get productive diagnostic results (Roussos, Templin, & Henson, 2007).

Finally, number of attributes can impact the attribute definition. As mentioned earlier, the attribute definition process involves the specification and definition of the attributes. The number of attributes then somewhat influences the level of specificity that researchers may choose to delineate and statistically estimate the attributes involved in the items. In other words, researchers need to justify their decisions of defining attributes at a finer-grained size or collapsing attributes at a broader sense, given practical considerations (e.g., computer program capacity, testing time). To illustrate, researchers might choose to consider a small number of attributes to mesure a specific construct, such as prorpotional reasoning. In contrast, they are interested in assessing student learning of broad content, they might decide to define attributes in a much broader sense.

Statistical Procedures

In this section, we describe the four critical statistical procedures in CDM: model selection, model estimation, model parameters interpretation, and Q-matrix validation.

Choosing the Right Model

It is important to run CDMs with a model that matches the construct attribute structure, particularly whether the attributes can be compensated or not. For example, item 2+3-1=?, which is shown in Rupp et al. (2008), requires both addition and subtraction skills. If one of the required attributes is lacking, it is impossible to answer this item correctly. Therefore, a model with a noncompensatory rule, such as DINA or G-DINA as opposed to DINO, should be used,

Several other factors should to be considered when selecting the model. For detailed review, see de la Torre, Hong, and Deng (2010) and Rojas, de la Torre, and Olea (2012). The first of which is whether the true model is known or not (always not known in practice; may be reasonably surmised). When the true model is known, using that model for estimation would be the suitable solution. But if the true model is unknown, the results from Rojas et al. (2012) can be used as a guide. When the underlying model is not known and the sample size is small, use of the G-DINA model resulted in a higher percentage of individuals being properly classified in each attribute than did the DINA, DINO, or A-CDM. When the sample size is small, the true model can provide better item parameter estimates than G-DINA, although G-DINA may be adequate in some cases. When the sample size is large, G-DINA can estimate item parameters in the middle and also yield close to the optimal estimates. When the true model is unknown, the attribute classification accuracy (ACA) in G-DINA is the best. When the test length is large, ACA from G-DINA is accurate, although the item parameters are not stable.

The second consideration is sample size. de la Torre et al. (2010) studied factors affecting the item parameter estimation and classification accuracy of the DINA model. They found that a sample size of 1,000 is sufficient to provide accurate (in terms of bias) parameter estimates, but the precision (in terms of SD) is improved as the sample size increased. Rojas et al. (2012) found that G-DINA is quite good (in terms of classification rates), compared to DINA, DINO, and A-CDM, when the sample size is small (e.g., 100 or 200), although this does not hold true in all scenarios. For example, the DINO model can provide a proportion of correctly classified individual attributes given the scenario that the true model is DINO, item quality is low, and sample size is small. Moreover, a sample size of 400 can provide relatively good results, compared to a larger sample size.

Model Estimation

As describe in an earlier section, it is important to know what kind of model estimation is appropriate for the model. For example, von Davier & Yamamoto (2004) used EM instead of MCMC because they said that EM works well for their models of interest, has much lower computational cost, can apply the IRT toolbox of methods (e.g., weighting, restrictions, and fitting), and can reuse available script languages and libraries efficiently.

Moreover, different models used different methods and software for analyses. While some models are used with a variety of software (e.g., the DINA model can be analyzed by Ox console, and CDM in R), some models need specific software and require research licenses (e.g., Full NC-RUM and Reduced NC-RUM in the Arpeggio program). Hence, researchers need to consider the model, the availability of necessary software, and in some cases, access to software expertise.

In practice, model estimation is embedded in the program mechanism. To run most programs, researchers need to provide the Q-matrix as an input of the program along with the data set in a format that is readable by the program. They also need to write the syntax. However, some programs provide easily modified usable codes. For example, the codes for DINA and G-DINA in Ox require only changing the numbers of the sample size, attributes, and items, and also the name of the Q-matrix and response data files (in space or tab-delimited format). Given these inputs, Ox can run the model estimation.

Interpretation of Model Parameters

After running the model and getting the results, all model parameters should be evaluated and interpreted. As stated by Roussos et al. (2010), “The estimates for the ability distribution and item parameters should be evaluated for internal consistency, reasonability, and concurrence with

substantive expectations." Basically, the goal of CDM is to make inferences about the attribute vector α_i or attribute pattern (de la Torre, 2014). Thus, student mastery should be reported. If a skill was much harder or easier than the expected, based on the cognitive model or content, the researcher may need to reconsider the Q-matrix, the item difficulty, or even the attribute definition.

In addition, the item parameters indicating how well the item distinguishes between masters and nonmasters need to be evaluated. Item parameters can guide the specification of the Q-matrix. Specifically, Roussos et al. (2010) indicate that, for an assigned skill, "changing the corresponding Q-matrix entry to 0 [results] in one less item parameter being estimated." Moreover, the decisions about the Q-matrix should be based not only on expert judgment, but also on fit statistics and validity studies.

To illustrate, the outputs of DINA and G-DINA include estimates of item parameters with the corresponding standard errors, posterior distribution of the attributes, examinee attribute classification, and item and test-level fit statistics. The estimates of item parameters with the corresponding standard errors in DINA and G-DINA are different and require different interpretations. The interpretation of the DINA model is straightforward because there are only two parameter estimates for each item: the guessing and slip parameters. The numbers of parameter estimates for each item in the G-DINA model depends on the attributes required for each item. Therefore, it requires interpretation of each attribute combination as well, and this information demonstrates the probability of success associated with each attribute pattern. The posterior distribution of the attributes shows the estimated proportion of each attribute pattern in the population. The examinee attribute classification informs the profile of individual examinees, specifically, which attributes he/she mastered. From this output, we can know which and how many examinees are in the same group or have the same attribute pattern, and which attribute pattern should be provided for remediation or improvement. The item-level and test-level fit statistics provide the absolute and relative fit statistics for model fit evaluation. The item-level fit statistics show the expected values of the item statistics (i.e., proportion correct, correlation, and log-odds ratio). These absolute fit statistics indicate how well the model fits the data; if it fits well, the results should be small (de la Torre, 2014). The test-level fit statistics, -2LL, AIC, and BIC, are used for the relative fit evaluation, which is important for comparing models and choosing the best fitting model. Again, the smaller values of the absolute fit statistics, the better the fit of the model.

Q-matrix Validation

Various validation methods can be used in CDM such as DIF analysis and comparison between pretest and posttest scores, as mentioned earlier in the CDM steps section. In this section, we focus on Q-matrix validation. The Q-matrix is normally developed based on expert judgment, which can be subjective and involve possible misspecifications. Misspecifications of the Q-matrix impact the parameter estimation and classification accuracy (de la Torre, 2014). For example, if attributes are mistakenly omitted from the Q-matrix, it impacts the estimated slipping parameter (Rupp & Templin, 2008). Thus, it is very important to validate the Q-matrix.

Several methods have been developed for validating the Q-matrix. One is the sequential EM-based δ-method for Q-matrix validation developed by de la Torre (2008) for selecting the optimal q vectors to improve model-data fit. This method is based solely on the available data. In the DINA model, this method can both identify and correct misspecified q vectors, and retain those that were appropriately specified simultaneously. The other method, index ς^2 (de la Torre & Chiu, 2015) can be used without assuming the specific CDMs involved, only that they are subsumed by the G-DINA model, which is quite general. So ς^2 is a generalization of the discrimination index specifically for the DINA model proposed by de la Torre (2008) and is used to give the model more applicability and generality.

In sum, CDM is a model that can be used to estimate examinees' latent cognitive constructs to provide individual students' learning profiles, guide teachers' instruction, and provide the items' psychometric properties. Various models have different characterisitcs, strengths, and weaknesses. In

short, the DINA model is parsimonious and very restrictive, whereas the G-DINA is the most flexible but involves more parameters, and requires a longer estimation time and a larger sample size.

Generally, there are related multiple steps involved in CDM such as attribute definition, validation, model fitting, and model selection. While other steps are highly statistics-based, the attribute definition is based on conceptual work that requires to deliberately clarify the assessment purposes and specify the analysis framework in order to guide the analysis. Also, the output interpretation is a refelective work that should relate back to the attribute definition. Specifically, the interpretation should determine whether the outputs make sense or not based on the definition framework.

The learning/cognitive sciences underly the attribute definition. Although CDM is very useful for assessment data analysis, it is more beneficial if the assessment design is based on a learning/cognitive sciences theory. A well-defined construct and items based on available cognitive and learning theory are needed to guide attribute definition. Thus, the researcher needs to consider several things when implementing CDM, specifically, cognitive and learning theory and statistical procedures. With those considerations, CDM can yield highly useful benefits for diagnostic purposes.

References

Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013) Incorporating student covariates in cognitive diagnosis models. *Journal of Classification, 30,* 195-224.

Broaddus, A., & Shaftel, J. (2012). Cognitive diagnostic assessment - informing responses and interventions. Retrieved from https://cete.ku.edu/sites/cete.drupal.ku.edu/files/docs/Presentations/2012_05_Broaddus_Shaftel%20CDA-RtI.pdf.

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., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633–665.

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

_____. (2009b). DINA model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.

_____. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199.

_____. (2014). Cognitive diagnosis modeling: A general framework approach. Retrieved from http://rci.rutgers.edu/~jdelator/Download/handout.pdf.

de la Torre, J., & Chiu, C. Y. (2015). General empirical method of Q-Matrix validation. *Psychometrika*. Advance online publication. doi:10.1007/s11336-015-9467-8.

de la Torre, J., & Douglas, J. (2004). A higher-order latent trait model for cognitive diagnosis. *Psychometrika*, *69*, 333–353.

_____. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595-624.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*, 227-249.

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*(4), 355-373.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*, 89-97.

De Carlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8-26.

Di Bello, L.S., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R Rao & S. Sinharay (Eds.) *Handbook of Statistics*, 26, (pp. 979-1030). Amsterdam: Elsevier.

Di Bello, L.S., & Stout, W. (2003). Student profile scoring for formative assessment. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *Proceedings of the International Meeting of the Psychometric Society IMPS2001* (pp. 81-92). Osaka, Japan, July 15–19, 2001.

Di Bello, L.S., Stout, W., & Roussos, L.A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Mahwah, NJ: Erlbaum.

Feasel, K., Henson, R., & Jones, L. (2004). *Analysis of the Gambling Research Instrument (GRI)*. Unpublished manuscript.

Gu, Z. (2011). *Maximizing the Potential of Multiple-Choice Items for Cognitive Diagnostic Assessment* (Unpublished doctoral dissertation). University of Toronto, Ontario, Canada.

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. A. (2002). *Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. (Doctoral Dissertation) Urbana-Champaign: University of Illinois.

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

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnosis 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. 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.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Unpublished manuscript. Prepared for the Committee on the Foundations of Assessment, National Research Council, November 30, 1999. Retrieved from http://www.stat.cmu.edu/~brian/nrc/cfa/documents/final.pdf.

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

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144-177.

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,* 3-16.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

Lesieur, H. R., & Blume, S. B. (1987). The South Oaks Gambling Screen (SOGS): A new instrument for the identification of pathological gamblers. *American Journal of Psychiatry, 144*(9), 1184-1188.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement. 33,* 579-598

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 197-212.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: IEA.

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *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., DiBello, L., Henson, R., Jang, E. E., & Templin, J. (2010). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. E. Embretson & J. Roberts (Eds.), *New directions in psychological measurement with model-based approaches* (pp. 35-69). Washington, DC: American Psychological Association.

Roussos, L., Templin, J., & Henson, R. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*(4), 293-311.

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

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

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

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

Templin, J., & Henson, R. A. (2005). *The random effects reparametrized unified model: A model for joint estimation of discrete skills and continuous ability*. Unpublished manuscript.

_____. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Templin, J., Henson, R., Rupp, A., Jang, E., & Ahmed, M. (2008, March).*Cognitive diagnosis models for nominal response data.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, NY.

Templin, J., Henson, R., Templin, S., & Roussos, L. (2008). Robustness of unidimensional hierarchical modeling of discrete attribute association in cognitive diagnosis models. *Applied Psychological Measurement, 32,* 559-574.

Thummaphan, P., Dong, D., & Li, M. (2015, August). *The use of cognitive diagnosis modeling analysis of word problem solving*. Paper presented in the 12th International Postgraduate Research Colloquium (IPRC), Bangkok, Thailand.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (RR-05-16). Princeton, NJ: Educational Testing Service.

_____. (2014), *The Log-Linear Cognitive Diagnostic Model (LCDM) as a Special Case of the General Diagnostic Model (GDM)*. ETS Research Report Series. 2014(2), 1–13, Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ets2.12043/abstract

von Davier, M., & Yamamoto, K. (2004, October). *A class of models for cognitive diagnosis*. Paper presented at the 4th Spearman Conference, Philadelphia, PA.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data (Research Report No. RR-06–08)*. Princeton, NJ: Educational Testing Service.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment (Research Report No. RR-03–32)*. Princeton, NJ: Educational Testing Service.