COGNITIVE DIAGNOSTIC MODEL COMPARISONS

A Dissertation
Presented to
The Academic Faculty

By

Yeongyu Lim

In Partial Fulfillment
Of the Reqiarements for the degree
Doctor of Philosophy in Psychology

Georgia Institute of Technology

May, 2015

Cognitive Diagnostic Model Comparisons

Approved by:

Dr. Susan E. Embretson, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Davood Tofighi
School of Psychology
Georgia Institute of Technology

Dr. Rick Thomas
School of Psychology
Georgia Institute of Technology

Dr. Charles Parsons
Scheller College of Business
Georgia Institute of Technology

Dr. Jonathan Templin
Department of Psychology and
Research in Education
The University of Kansas

CHAPTER 1

INTRODUCTION OF COGNITIVE DIAGNOSTIC ASSESSMENT (CDA)

1.1 Definition and importance of CDA

The main purpose of the traditional psychometric theories is to assign numbers systematically to psychological variables. In order to accurately measure unobservable psychological variables, psychometricians have developed two major measurement theories. One is classical test theory (CTT) and the other one is item response theory (IRT). IRT is a current mainstream in the field of psychometrics. In IRT, the relationship between an examinee' responses on test items and a latent variable is specified by a mathematical function, and the test result for the examinee produces only a single score as a measure of an underlying latent variable. These test results are useful for ranking and comparing examinees, but do not provide valuable information about the performance on specific domains within the latent variable.

Psychometricians also have developed many statistical techniques (e.g., factor analysis, structural equation modeling and cluster analysis). These methods allow various psychometric models to be fitted to different data sets. However, they tend to focus on "metrics" rather than "psychology" and have a limitation to understand fundamental psychological nature of the knowledge and skills measured in a test (Leighton & Gierl, 2007a). Anastasi (1967) argued that "those psychologists specializing in psychometircs have been devoting more and more of their efforts to refining techniques of test construction, while losing sight of the behavior they set out to measure" (p. 297). Mislevy (1993) also described such situation as "the application of the 20th century statistics to the 19th century psychology" (p19). As a solution of this problem, Snow and

Lohman (1989) suggested to integrate cognitive psychology into psychometrics. Embretson (1983) also emphasized the application of cognitive psychology to measurement theory and introduced contemporary notions of construct validation. Awareness of the limitations of IRT and other psychometric methods promoted a new methodology that can provide specific diagnostic information to understand examinee's cognitive processes.

In addition to the need of change in the field of psychometrics, current mission in education in the United States known as the "No Child Left Behind Act of 2001 (NCLB)" increased the importance of diagnostic tests to be informative about individual students' cognitive strengths and weaknesses. Many states have struggled with developing large-scale assessments to evaluate the progress of schools toward the achievement of educational standards. In order to accommodate such need of stakeholders, psychometricians have developed a new theoretical framework, cognitive diagnostic assessment (CDA), in collaboration with cognitive psychologists.

In general, assessment refers to the process of documenting, usually in measurable terms, knowledge, skills, attitudes and beliefs (Mislevy, Steinberg, & Almond, 2003). In educational perspective, assessments should promote students' learning and the service of instruction (Griffin and Nix, 1991; Snow, 1989). CDA is a theoretical framework which purposes on diagnosis of examinees' skill profiles. CDA is designed to measure specific knowledge structures and cognitive processing skills in a given domain. Researchers in educational measurement and cognitive psychology have been dedicated to develop tests to provide useful information regarding students' cognitive strengths and weaknesses in specific knowledge structure and processing skills.

Students can improve the skills that they do not possess by understanding the diagnostic feedback. This cognitive diagnostic feedback also helps instructors in their teaching processes. Therefore, CDA is an innovative methodology that can be an appropriate theoretical foundation for standards-based assessments. The CDA approach is based on the cognitive view of learning. In the cognitive view, knowledge is acquired through a systematic processing of information, and learners understand concepts through reasoning and. They are supposed to use cognitive and meta-cognitive strategies for problem solving and transfer new knowledge to other tasks. This view of learning has applied to various assessment situations such as mathematical tests and language tests.

1.2 Principled design of diagnostic assessments

Three important frameworks for principled design of diagnostic assessments will be discussed in the following section.

1.2.1 The cognitive design system (CDS) approach

CDA requires the test based upon a substantive theory of the construct that describes the cognitive processes. Item or task characteristics that are intended to elicit the cognitive processes should be clearly specified in a test. Embretson (1994; 1998) suggested the CDS approach to incorporate cognitive theory into the development of ability tests. Two different frameworks are included in the CDS approach: a conceptual and a procedural framework (Embretson, 1995). First, the conceptual framework was motivated to expand the traditional conceptualization of construct validity represented by Cronbach and Meehl's (1955) and Bechtoldt's (1959) view. Embretson’s conceptualization of construct validity includes two distinguished issues, namely

construct representation and nomothetic span (Embretson, 1983).

"Construct representation refers to the relative dependence of task responses on the processes, strategies, and knowledge stores that are involved in performance" (Embretson, 1983, p. 180).That is, this aspect of construct validity is related to the methods of cognitive psychology to explain cognitive processes that are involved in solving psychometric items (Embretson, 1985). In contrast, nomothetic span refers to "the strength and nature of the relationship of the construct that is measured by the assessment to other theoretically relevant constructs." (Rupp, Henson, & Templin, 2010) This relationship is investigated by the correlations of a test score with other measures.

Next, procedural framework is to concern item parameter estimation as well as item selection and development in relation to cognitive theories (Embretson, 1994). The principled assessment design processes in the CDS suggested by Embretson (1998) are roughly illustrated in Table 1.

Table 1 Processes in Cognitive Diagnostic System

Specify general goals of measurement
Construct representation (meaning)
Nomothetic span (significant)
Identify design features in task domain
Task-general features (mode, format, conditions)
Task-specific features
Develop a cognitive model
Review theories
Select or develop model for psychometric domain
Revise model
Test model
Evaluate cognitive model for psychometric potential
Evaluate cognitive model plausibility on current test

Continued

Evaluate impact of complexity factors on psychometric properties Anticipate properties of new test Specify item distributions on cognitive complexity Distribution of item complexity parameters Distribution of item features Generate items to fit specifications Artificial intelligence? Evaluate cognitive and psychometric properties for revised test domain Estimate component latent trait model parameters Evaluate plausibility of cognitive model Evaluate impact of complexity factors on psychometric properties Evaluate plausibility of the psychometric model Calibrate final item parameters and ability distributions Psychometric evaluation Measuring processing abilities Banks items by cognitive processing demands Assemble test forms to represent specifications Fixed content test Adaptive test Validation: Strong program of hypothesis testing

1.2.2 The evidence centered design (ECD) framework

Evidence-centered assessment design (ECD) is an approach to constructing educational assessments that incorporate evidentiary arguments (Mislevy, Steinberg, & Almond, 2003). The tasks in the ECD framework are designed to provide evidence about targeted knowledge and skill (Mislevy, Almond, & Lukas, 2004). The design of such tasks is achieved by describing the evidentiary arguments that underlies an assessment. The terminology introduced by Toulmin (1958) was adapted to represent the structure of the evidentiary arguments. Figure 1 shows an example of Toulmin's (1958) schema. The

claim (C) is a proposition that is supported by a data. This is a statement about examinee's proficiency in the ECD framework (Mislevy, et. al., 2003). The data (D) are examinee's behavior. A warrant (W) is expected outcomes of respondents having the ability in certain conditions, and it requires backing (B) in the form of theories, research, data, or experience (Mislevy, et. al., 2003). Alternative (A) for the observed data (D) should qualify the inference, and alternative hypotheses also can be supported or weakened by rebuttal (R) data.

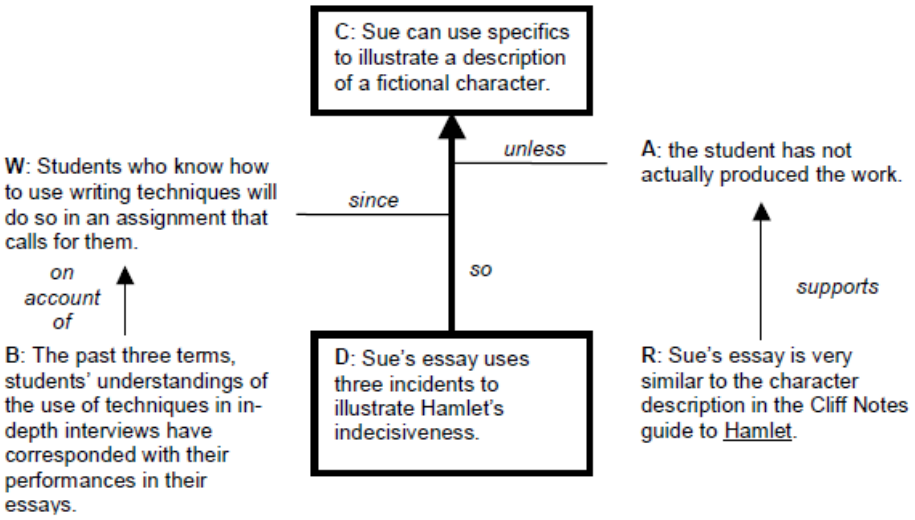

Figure 1. An Example of the Assessment Argument Depicted as a Toulmin Diagram

The ECD process consists of 4 steps (Mislevy, et. al., 2003). To begin with, 'Domain Analysis' is to concern substantive information about the domain. For example, designing math test includes analysis of theories of mathematical proficiency.

The next stage of design is 'domain modeling' consisting of three paradigms. First, proficiency paradigms indicate potential claims about examinees' abilities or competencies. Second, evidence paradigms deal with observable features in examinees'

performances as evidence about these proficiencies. Third, task paradigms are the kinds of situations that provide an opportunity for examinees to demonstrate evidence of their proficiencies the evidence.

In the third stage, Conceptual Assessment Framework (CAF) includes five design components; the student models, the task models, the evidence models, the assembly model, and the presentation model. These CAF models lay out the blueprint for an assessment. First, the student model is a statistical characterization of examinees. It specifies the variables or aspects of learning that we wish to characterize students (Mislevy, et. al., 2003). Since many aspects of knowledge or skills that we want to assess are not directly observable, the student model provides a probabilistic model for making claims about the underlying traits (Rupp et.al., 2010). Almond and Mislevy (1999) pointed out that a student model can be viewed as a fragment of a Bayesian inference network, or Bayes net that takes the form of probabilistic distributions for student-model variables, and for observable variables conditional on student-model variables. Second, the task model specifies characteristics of tasks that students do in the test. A task model provides information about conditions or formats in which what the student says, does, or produces. Task-model variables also contain characteristics of stimulus material, instructions, help, tools, and so on (Mislevy, et. al., 2003). Third, the evidence model specifies the evaluation rules for scoring test tasks (the evaluation component; task scoring) and a mechanism used to accumulate data across tasks to update the student model variables in examinees' scoring models (the measurement component; test scoring). Fourth, the assembly model describes how the student models, evidence models, and task models are linked together to form a particular assessment (Mislevy, et. al.,

2004). It also describes the strategy used for combining tasks that are selected for gathering evidence from students (Mislevy, et. al., 2003). The assembly model concerns statistical characteristics of items such as their difficulty and non-statistical considerations such as the content of reading passages, timing, sentence complexity, and many other task features. The idea of item banks in the assemble model is similar to the idea of automatic item generation within the CDS frameworks (Rupp et.al., 2010). Last, the presentation model specifies how tasks are presented and interacted with students, and what types of format are used (e.g., paper and pencil test vs. computer adaptive test). The presentation model also describes how different parts of the assessment (student, evidence, and task models) will actually operate within a particular delivery environment.

As the last stage, operational assessment is a model that deals with people's relation to assessment as something experienced (Mislevy, et. al., 2003). Four-process delivery system describes how an operational assessment is functioning (Mislevy, et. al., 2003). First, presentation indicates something presented, interaction between the information and the test-taker, and capturing test-takers' responses. Second, response scoring is to evaluate observations. The third process is summarizing the scores across several responses. Last, activity selection is making decision about what might be useful to do next (Mislevy, et. al., 2003).

In conclusion, ECD framework provides a framework for clear articulation of what is being measured to establish test validity. Glas (2003) indicated the ECD framework is a comprehensive model for the process of educational measurement. He also stated that "the ECD model can stimulate a lot of new developmental work in the field of psychometrics." (Glas, 2003)

1.2.3 Assessment engineering

Assessment engineering (AE) is to an innovative approach to measurement with replicable, scalable solutions for assessment design, item writing, test assembly, and psychometrics (Luecht, 2008). In AE framework, traditional test blueprints and related specifications were replaced by evidence models and cognitive task models. AE directs the test development as well as the analysis, scoring, and reporting of assessment results using engineering-based principles. The concept of AE is based on the idea described by Drasgow, Luecht, and Bennett (2006): "Our vision of a 21st-century testing program capitalizes on modern technology and takes advantage of recent innovations in testing. Using an analogy from engineering, we envision a modern testing program as an integrated system of systems" (p. 471).

AE includes the following five processes. The first stage is to define the constructs driven from cognitive models of task performance. The cognitive models underlying examinees' task performance is designed to specify the knowledge requirements and processing skills (Zhou, 2009). In AE framework, cognitive models guide item development, rather than content blueprints. The cognitive models can be made by adapting a variety of procedures (e.g., judgmental and logical analyses, generalizability studies, and analyses of group differences; Messick, 1989). The verbal report method is a broadly used way to generate a cognitive model. In this method, first of all, tasks are administered to a group of examinees who represent the intended population. Next, they are forced to think aloud when they solve items, and then protocol or verbal analysis is conducted with the corresponding verbal data (Leighton, 2004; Leighton & Gierl, 2007a).

The construct-based validation is not a new concept. Messick stated that “A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed presumably because they are tied to explicit assessed or implicit objectives of instruction or otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors?” (Messick, 1994, p. 16)

Construct mapping is to visually represent a cognitive model (Wilson, 2005). An important feature of the construct map is that there is a coherent and substantive definition for the content of the construct (Wilson, 2005). Another characteristic of the construct map is that the construct is composed of an underlying continuum, and this can be manifest by ordering the respondents and/or item responses (Wilson, 2005). Therefore, construct maps can be viewed as an ordered performance expectations at various levels on a scale (Luecht, 2008). See Figure 2 for details. The construct of interest, X, is composed of an ordered set of latent classes

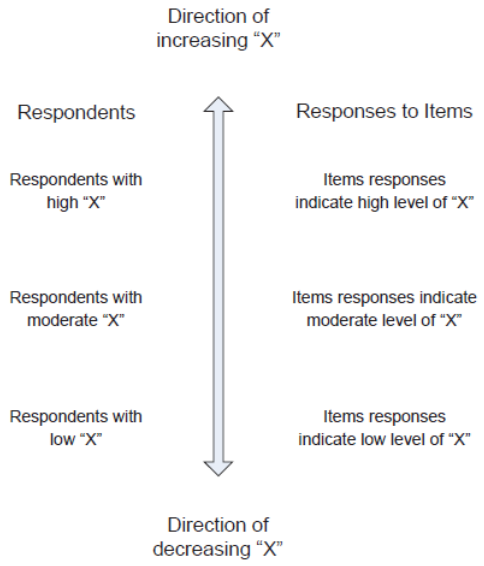

Figure 2. A Generic Construct Map for Construct “X”

In the second stage, evidence models as well as cognitive models are developed to specify particular proficiency claims. Evidences for the claims should be tangible actions, responses, and/or products (Luecht, 2008). Components of an evidence model include valid settings or contexts, the plausible range of challenges for the target population, relevant actions that could lead to a solution, legitimate auxiliary resources, aids, tools, etc. that can be used to solve the problem, concrete exemplar products of successful performance (Luecht, 2008).

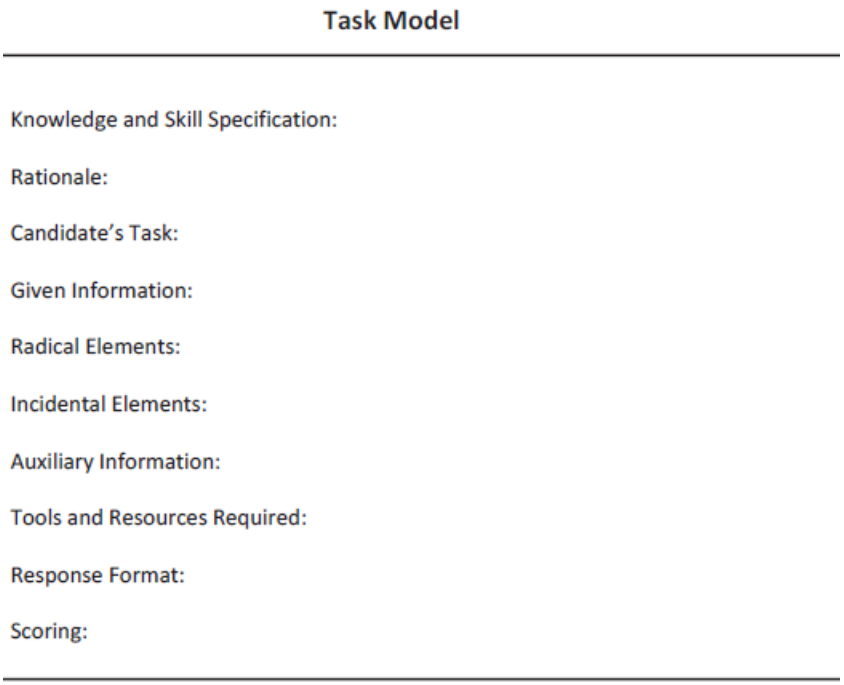

Figure 3. A Generic Template of a Task Model

In the third stage, item models and task models should be developed to produce replicable assessment tasks. A task model can be viewed as a new way to blueprint. It

describes knowledge and skills, descriptions of key features (e.g., objects and their properties, variables for difficulty variation) of the task, specifications of task representation material and any required condition, and classifications of response actions returned for scoring (Zhou, 2009). Zhou (2009) showed a generic template for a task model (See Figure 3). The knowledge and skills measured by tasks and any required conditions for the objects are provided from these models. Task modeling differ from traditional test development approach because it provides theoretical backing for item development by regulating features of assessment tasks such as stimulus elements and conditions (Zhou, 2009).

On the other hand, item modeling provides operational foundation for efficient and accurate item generation. Explicit item models are intended to control and manipulate both the content and difficulty of the items (Zhou, 2009). Items produced by each item model are intended to have high psychometric standard consistently.

An item model serves as an explicit representation of the parts listed in a corresponding task model. An item model consists of the stem, the options, key, and oftentimes auxiliary information. The stem formulates context, content, and the question that the examinee is required to answer, and the options contain the alternative answers with one correct answer and incorrect options or distracters. When dealing with an open-ended or constructed-response item model, options are not required. The key specifies the correct answer for a multiple choice item model or lists of criteria for an open-ended item model. Auxiliary information includes any additional materials required to generate an item in the stem or option, such as texts, images, tables, or diagrams. Elements in the stem and options are denoted as strings (S) which are non-numeric variables and integers

(I) which are numeric variables. Measurement specialists can generate large numbers of items, at similar or different difficulty levels, for each model by systematically manipulating these elements (Zhou, 2009). Figure 1.5 shows an example of an item model.

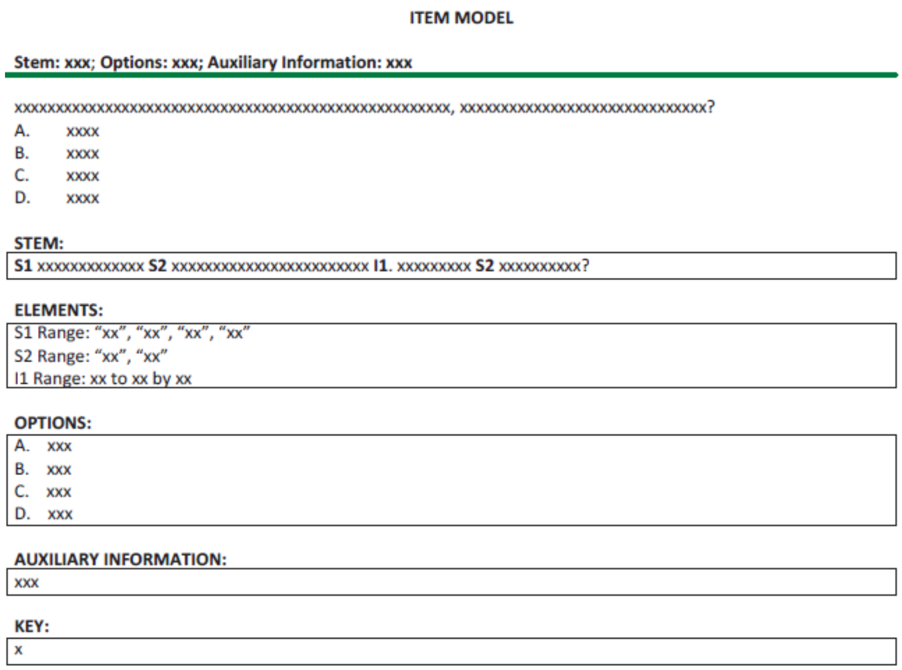

Figure 4. A Generic Template of an Item Model

Computer-based procedures for test assemblies have been developed to select items efficiently (van der Linden, 1998; van der Linden, 2005). These procedures are referred to automated test assembly in the AE framework. This is the fourth stage of AE processes. With computer algorithms, it is possible to manage and construct tests in a much more efficient way than traditional manual assembly procedures (Zhou, 2009). Automated test assembly procedures consist of three steps. The first step is to define item bank by finding the characteristics of each item such as psychometric features, content

categories, and any other relevant item features. The second step is to develop a mathematical model that incorporates all psychometric and content specifications of the test. In the next step, with the model, every possible solution is evaluated until the optimal or best possible combination of available items is achieved. Multiple test forms can be created from these procedures.

In the fifth stage, psychometric models are employed in a confirmatory manner to assess the model-data fit relative to the intended underlying structure of the constructs or traits the test is design to measure (Zhou, 2009). In traditional approach to test development, a model is explored through a validation study after a test is administered and response data is collected. However, in AE framework, a model is statistically confirmed by assessing the consistency between expected and observed responses to ensure cognitive principles. The outcomes from the model-data fit analyses allow test developers to modify the cognitive and item models. This confirmatory method provides a direct connection between cognitive theory and educational measurement (Luecht, Gierl, Tan, & Huff, 2006).

CHAPTER 2

THEORETICAL FOUNDATIONS OF COGNITIVE DIAGNOSTIC MODELS

2.1 Overview

The union of cognitive psychology and psychometrics led to cognitively based psychometric models since 1980's. Furthermore, with the increased number of large-scale assessments, it was necessary to develop appropriate psychometric tools for interpretation of the test results. For these reasons, researchers have proposed many psychometric models to implement CDA over the last two or three decades. Cognitive diagnostic models (CDM) can be used to demonstrate how well standards-based assessments classify students' level of proficiency.

In educational measurement, a cognitive model refers to a, "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills examinees at different levels of learning have acquired and to facilitate the explanation and prediction of examinees' performance" (Leighton & Gierl, 2007a, p. 6). Components of a cognitive model are used as attributes to generate diagnostic inferences underlying test performance. CDMs refer to psychometric models based on a theory of response processing grounded in applied cognitive psychology (Rupp & Templin, 2008). Attributes are defined as a description of knowledge, different procedures, or skills needed by an examinee to solve a given test item. CDMs enable multiple criterion-referenced interpretations because those models contain relatively fine-grained proficiency dimensions (e.g., multiple attributes), rather than course-grained global proficiency variables. In classification-based decision making situation (e.g., placement, admission, or certification), examinees are classified into non-mastery or

mastery on a specific domain. In addition to the course screening purpose of diagnostic tests, diagnostic results analyzed with CDMs can be also used to provide more detailed information on a particular area, and make it possible remedial interventions or treatments.

2.2 Characteristics of CDMs

Characteristics of CDMs can be shown by comparing with other latent variable models. The following four characteristics of CDMs describe similarities and differences between the models. First, CDM has confirmatory nature because its analysis involves a hypothesis testing to confirm the relationship between the items and the attributes (Rupp, et al., 2010). Most statistical model structures of CDMs involve pre-determined Q-matrix in which the loading structure is specified (Tatsuoka, 1983). The rows of the Q-matrix pertain to the items, whereas the columns the attributes. The 1s in the *jth* row of the Q-matrix identifies the attributes required for item *j*.

Second, CDMs, IRT and FA models have similarities in the aspect of the observed response variables. Like traditional IRT models, CDMs deal with both dichotomous and polytomous response data (Rupp, & Templin, 2008). Even though most educational achievement assessments are dichotomously scored (e.g., 1 for correct answer and 0 for in correct answer), ordered response scales or Likert scales also used in CDMs. There are some CDMs to handle polytomous data (e.g., the reduced non-compensatory reparameterized unified model; reduced NC-RUM). Different types of CDMs according to the response variables will be specifically discussed in next section.

Third characteristic of CDMs is related with interpretation of criterion-referenced tests. Rupp and Templin (2008) pointed out that multiple criterion-referenced interpretations are possible in CDMs. In contrast, most of multidimensional FA or IRT models with continuous latent variables allow for multiple norm-referenced interpretations. Interestingly, some of recently developed large-scale, standards-based assessments have multiple cut-points (e.g., “below standard,” meets standard,” “exceeds standard”) for scales estimated with traditional FA or IRT models (Rupp & Templin, 2008). However, the criterions in those classification methods are set consensually (i.e., largely model-external) while the cut scores in CDMs are statistically set to maximize the reliable reparation of respondents (i.e., model-internal) (Rupp, et al., 2010). Therefore, those classifications in standardized achievement assessments for accountability purposes are based on decision-making processes by human judges rather than the results of the application of a psychometric model (Rupp, et al., 2010).

Last, CDM has multidimensional nature like multidimensional IRT models or factor analysis (FA) models. However, the definitional grain size of the constructs differ each other. While constructs in multidimensional FA models are broadly defined, the latent variables (i. e., attributes) in CDMs are more narrowly defined (Rupp, et al., 2010). This difference leads to different complexity of the loading structure between CDMs and other models. Multidimensional IRT and confirmatory FA (CFA) have a simple loading structure that each item only loads on one dimension (Rupp, & Templin, 2008). In contrast, loading structures in CDMs are complex because multiple component skills are needed for each item as reflected in the Q-matrix. This type of loading structure is called “within-item multidimensionality” while simple loading structure corresponds to

"between-item multidimensionality" (Adams, Wilson, & Wang, 1997). Rupp, et al. (2010) indicated that even though within-item multidimensionality can be handled within the context of multidimensional IRT and CFA models, they provide only a continuous multidimensional profile, not a discrete multidimensional profile as in CDMs. In addition, multidimensional latent variables in FA or IRT are different constructs or different aspects of the same construct rather than elementary mental components and their interaction (Rupp, & Templin, 2008).

2.3 A taxonomy of CDMs

CDMs can be classified by considering the following two characteristics: (1) the scale type of the latent variables (i.e., latent class models vs. latent trait models), and (2) the compensatory or non-compensatory combination of the latent attribute variables. A latent class model classifies examinees into categories on a set of skills (e.g., mastery vs. nonmastery) by providing attribute mastery patterns or mastery probabilities. On the other hand, a latent trait model estimates examinee's ability on a continuous scale for each attribute. Thus, this family of CDMs can be viewed as an extension of unidimensional IRT models. However, they are non-compensatory models while traditional IRT models compensatory latent variable models (Rupp & Templin, 2008).

CDMs are also dived into non-compensatory models and compensatory models. Non-compensatory models include conjunctive and disjunctive models. A conjunctive condensation rule refers to a formula that states all attributes measured by an item need to be mastered for respondents to provide a correct response (Rupp., et. al., 2008). In non-compensatory models, one cannot "make up" for nonmastery of attributes by mastery of

other attributes (Henson, Templin, & Willse, 2009). This means that successful application of all the required skills is necessary for successful performance on a task. These conjunctive models are applicable for skill diagnosis where the solution of a task is broken down into a series of steps with conjunctive interaction rather than with compensatory interaction (Roussos, DiBello, Henson, Jang, & Templin, 2008). Conjunctive models are mostly used for mathematical tests which require all skills to perform successfully on an item (Tatsuoka, 1990). In a disjunctive case of models, successful performance of the item only requires that a subset (in some cases only one) of the possible strategies is successfully applied (DiBello, Roussos & Stout, 2007). Therefore, there is no additional benefit for having more than one attribute. Even if an examinee mastered only a subset of the required attributes on an item, performance on the item would be same as an individual mastered all of the required attributes (Henson. et. at., 2009). Disjunctive models are appropriate when multiple strategies exist to solve the item.

On the other hand, compensatory models allow an individual to “make up” for what is lacked in one skill by having mastered another. In other words, a high level of competence on one skill can compensate for a low level of competence on another skill to result in successful performance on a task. Make sure that compensatory models are not identical to disjunctive models. In the case of disjunctive models, mastering at least one of all required attributes can be allowed, but those attributes do not compensate each other. The compensatory models are appropriate for medical and psychological disorder diagnosis, where the presence of a certain symptom can compensate the presence or

absence of other symptoms (Rousoss et al., 2008). The list of the CDMS for the two categories is shown in Table 2.

Table 2 A Taxonomy of CDMs

	Noncompensatory	*Compensatory*
Latent Class Models	DINA (Haertel, 1989) NIDA (Junker & Sijtsma, 2001) DINO (Templin & Henson, 2006) NIDO (Templin, 2006) BIN (Levy & Mislevy, 2004) Unified Model (DiBello, Stout, & Roussos, 1993) & NC-RUM (Hartz, 2002) LCDM (Henson, Templin, & Willse, 2009) RSM (K. K. Tatsuoka, 1983) AHM (Leighton, Gierl, & Hunka, 2004)	BIN (Levy & Mislevy, 2004) C-RUM (Hartz, 2002) LCDM (Henson, et.al., 2009) GDM (von Davier, 2005a)
Latent Trait Models	LLTM (Fischer, 1973) MLTM (Embretson, 1980) GLTM (Embretson, 1984) MLTM-D (Embretson & Yang, 2008)	

Note: RSM = Rule-space method. AHM = Skill hierarchy method. BIN = Bayesian inference network. DINA = Deterministic inputs, noisy 'and' gate. LCDM = Loglinear

cognitive diagnosis model. DINO = Deterministic inputs, noisy ‘or’ gate. NIDA=Noisy inputs, deterministic ‘and’ gate. NIDO = Noisy inputs, deterministic ‘or’ gate. NC-RUM = Non- compensatory Reparametrized unified model /Fusion model. C-RUM = Compensatory RUM. NC-RUM = Non-compensatory RUM. GDM = General diagnostic model. LCDM = Loglinear cognitive diagnosis model. LLTM = Linear logistic test model. MLTM = Multidimensional latent trait model. MLTM-D = Multidimensional latent trait model for diagnosis. GLTM = General component latent trait model.

In the non-compensatory category, DINA and NIDA are conjunctive models while DINO and NIDO are disjunctive models. BIN appears in both columns of the table because BIN is a general modeling framework for representing different types of latent variable models (Rupp & Templin, 2008). General diagnostic model (GDM; von Davier, 2005) and loglinear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009) are general models that subsume other latent variable CDMs and provide great flexibility. LCDM is known as a model fit to all core models (i.e., DINA, DINO, NIDA, NIDO, RUM, C-RUM, and some applications of the GDM), excluding RSM and AHM (Henson, 2009). As a result, this model must be classified as a compensatory model as well as a non-compensatory model (Henson, 2009). Details of each model will be described in Chapter 3 and 4.

CHAPTER 3

RULE SPACE METHOD

3.1 Introduction of RSM

K. K. Tatsuoka is a pioneer in the development of rule space method (RSM; Tatsuoka, 1983; Tatsuoka, 1985; Tatsuoka, 1990; Tatsuoka, 1995; Tatsuoka, 2009; Tatsuoka, 1971). She introduced Q-matrix theory and rule space model to diagnose examinees' knowledge levels. The Q matrix is an incidence matrix of the attributes involved in each item (Tatsuoka, 1993). The scores in the Q matrix are binary and of order K x M, where K is the number of attributes and M is the number of items (Tatsuoka, 1990). The rows represent attributes (i.e., tasks, subtasks, cognitive processes and skills, etc.), and the columns represent items. The elements in the Q matrix are mostly scored by domain experts (e.g., teachers) or cognitive psychologists. The Q matrix can be viewed as the cognitive blueprint or specifications for the test, since the underlying cognitive tasks required in answering test items are specified in a Q-matrix (Gierl & Leighton, 2007b). In other words, the Q matrix is a cognitive model for test item performance hypothesized by cognitive experts in the domain (Tatsuoka, 2009). The Q matrix theory has been adapted for developing many other CDMs as well RSM.

Tatsuoka's rule-space model is a probabilistic model for classifying examinees' item responses into a set of attribute-mastery patterns associated with different cognitive skills. This cognitive diagnostic model is based on the item response theory (IRT) models to formulating the rule space. The rule space is a two-dimensional Cartesian coordinate system that maps knowledge states and observed response

patterns are mapped into a classification space by calculating ability parameters ($\hat{\theta}$) and measures of atypicalty (ζ). The first dimension corresponds to the ability or proficiency variable ($\hat{\theta}$). This dimension specifies variation in the response patterns that can be attributed to differences in examinee proficiency levels. The proficiency level can be evaluated through the ability estimation methods. The second dimension corresponds to the variable ζ (Tatsuoka, 1984, 1985) which measure the unusualness of item-response patterns. Figure 5 shows an example of the rule space with four different points. The details on the ability estimation methods and the ζ index will be illustrated later.

Figure 5. Example of the Rule Space

3.2 The Person-Fit Index, Zeta

The ζ index as one of the person-fit indexes was developed to detect aberrant response patterns caused by unusual behavior such as cheating or guessing (Tatsuoka, 1984). The ζ index has been used for cognitive error diagnoses in the context of rule space model (Tatsuoka, 1985). The purpose of defining the ζ index is to find a continuous variable or a function of θ that is orthogonal to the first axis θ, and then define a Cartesian product space $\{(\theta, \zeta)\}$ as a classification space of item response patterns in which an observed item response pattern will be classified into one of the predetermined knowledge states (Tatsuoka, 2009).

The ζ_i associated with a particular response vector X_i is calculated as follows:

$$\zeta_i = f(\theta_i, x_i)/[Var\{f(\theta_i, x_i)\}]^{1/2} \tag{3.1}$$

where $f(\theta_i, x_i) = -\sum_{j=1}^{n}(P_j(\theta_i) - T(\theta_i))x_{ij} + \sum_{j=1}^{n}(P_j(\theta_i)(P_j(\theta_i) - T(\theta_i))$,

$$Var\{f(\theta_i, x_i) = \sum_{j=1}^{n} P_j(\theta_i)Q_j(\theta_i)(P_j(\theta_i) - T(\theta_i))^2$$

and $T(\theta_i) = 1/n(\sum_{j=1,n} P_j(\theta_i))$.

$Q_j(\theta_i)$ is 1- $P_j(\theta_i)$, and the true score function. $T(\theta_i)$ is the average of n item response functions over all items. $P_j(\theta_i)$ is the probability of a correct response to the j^{th} item by the i^{th} examinee as determined from the assumed IRT model. The mathematical expression of $P_j(\theta_i)$ is given by

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}}$$

(3.2) where D is a scaling constant, a is the item discrimination parameter and the b_i is the item difficulty parameter (Birnbaum, 1968).

In the ζ index, the numerator is the conditional covariance of the two residual vectors **P**(θ) – **X, and P**(θ)- **T**(θ), and the denominator, is the conditional standard deviation of the numerator. The expectation of the numerator is zero (Tatsuoka, 1985). Note that $P(\theta_i) - X_i$ measures the deviation of the item response vector X_i from its expected value $P(\theta_i)$, and $P(\theta_i) - T(\theta_i)$ measures the deviation of the expected value of the response vector X_i from the overall average probability of a correct response at θ_i. An important property of ζ is that it reflects whether the response pattern conforms to a Guttman pattern or not. If a response pattern has zeros for the harder items and ones for easier items, then $f(\theta_i, x_i)$ has a negative value. That means that when a response pattern conforms to a Guttman scale defined by the ascending order of $P_j(\theta_i) - T(\theta_i)$. Thus, ζ indicates the extent to which a response vector X approximates the "Guttman vector" at θ. However, when a response pattern conforms to a reversed Guttman scale, ζ becomes a larger, positive number. Thus, ζ indicates how well respondents' patterns accord with the underlying IRT model (Sheehan & Kathleen, 1993). Another important property of ζ is that items having similar cognitive tacks have similar residual functions, $P_j(\theta_i) - T(\theta_i)$, because these items have similar estimated IRT parameter values (Tatsuoka, 1987).

3.3 Analysis of Knowledge States

Constructing a Q matrix is a critical part to identify knowledge states in the rule space analysis because it became a foundation for analyzing knowledge states as classification categories. In addition to the Q matrix, an attribute mastery pattern matrix should be constructed to identify all possible combinations among *K* attributes. Those theoretical response patterns are called as 'ideal response patterns'. Generating ideal response patterns is a critical part in the RSM analysis. Note that there are 2^k possible attribute mastery patterns. For example, when 10 attributes are identified, 2^{10} attribute mastery patterns correspond one-to-one to 2^{10} ideal response patterns. Ideal response patterns that are derived from the Q matrix and the attribute mastery pattern matrix by applying Boolean algebra are considered knowledge states by assuming a specific combination of mastered and nonmastered attributes (Tatsuoka, 1995). Thus, the ideal response patterns as the knowledge states represent mastery and non-mastery of specific combinations of attributes (Tatsuoka, 1993). If five attributes are involved to solve an item, 2^5 response patterns can account for an incorrect answer for the item. On the other hand, when all the attributes involved in an item have been mastered, the item is assumed to be answered correctly (Tatsuoka, 1993).

3.4 Rule Space Classification

After the observed item response vectors are projected onto the two-dimensional rule space, an admissibility criterion for each possible state is determined by the Mahalanobis distance (D^2) between the examinee's points in the rule space and the points associated each of the ideal response vectors (Fukunaga, 1972; Tatsuoka,

1971). The estimated Mahalanobis distance of X, a point in the rule space corresponding to the observed response pattern from the j knowledge state is

$$D_{xj}^2 = (X - R_j)' \sum^{-1} (X - R_j) \quad j = 1,..., j \tag{3.3}$$

where $R_j = \begin{vmatrix} \theta_{R_j} \\ \zeta_{R_j,X} \end{vmatrix}$ and $X = \begin{vmatrix} \theta_X \\ \zeta_{\theta_X X} \end{vmatrix}$

and the covariance matrix is $\Sigma = \begin{vmatrix} Var(\theta_X) & 0 \\ 0 & Var(\zeta_{R_j,X}) = 1 \end{vmatrix}$

Equation (3.3) can be simplified as follows:

$$D_{xj}^2 = (X - R_j)' \sum^{-1} (X - R_j) \tag{3.4}$$

$$= (\theta_X - \theta_{R_j}, \zeta_{\theta_X,X} - \zeta_{R_X,X}) \left\| \begin{matrix} 1/Var(\theta_{R_j}) & 0 \\ 0 & 1 \end{matrix} \right\| \begin{matrix} \theta_x - \theta_{R_j} \\ \zeta_{\theta_x X} - \xi_{R_j X} \end{matrix} \|$$

$$= (\theta_X - \theta_{R_j})^2 / I(\theta_{R_j}) + (\zeta_{\theta_X,X} - \zeta_{R_X,X})^2 / Var(\zeta_{R_X,X})$$

where $Var(\zeta_{R_X,X}) = 1$ and $I(\theta_{R_j})$ = item information = $1/Var(\theta_{R_j})$.

In the traditional classification, each examinee is forced to be classified into one of the predetermined knowledge states by finding small error probability as well as the closest Mahalanosis distance. This Bayes decision rule for classification is only applicable to the case when the number of groups is very few. To resolve these problems, Tatsuoka (2009) now classifies response pattern X into a group of several knowledge states located in the neighborhood of X, not a single knowledge state. Further, these close states are then used to compute attribute mastery probabilities, defined as “the probability of applying each attribute correctly to answer the items in a test” (Tatsuoka, 2009).

To compute the attribute mastery probabilities, obtaining posterior probabilities for each knowledge state should be preceded. Suppose there are L attribute mastery patterns associated with L closest knowledge states $\pi_1, \ldots, \pi_L$ and their posterior probabilities $q_1, \ldots, q_L$. The posterior probability of π_L is denoted by $q_l = p(\pi_l \mid X)$, for l = 1, …, L. $q_l = p(\pi_l \mid X)$ is calculated as follows:

$$q_l = p(\pi_l \mid X) = \frac{p_l p(X \mid \pi_1)}{\sum_{k=1}^{L} p_k p(X \mid \pi_k)} = \frac{p_l \exp(-.5D_l^2)}{\sum_{k=1}^{L} p_k \exp(-.5D_l^2)} \tag{3.5}$$

where D_l^2 is squared Mahalanobis distance for l = 1, …, L, and p_l is a prior probability on a given distribution (e.g., uniform or gamma distribution).

With L attribute mastery patterns with L closest distances and their posterior probabilities, it is possible to calculate the attribute mastery probabilities. The probability of attribute A_k for a given response pattern X is given by Equation (3.6):

$$p(A_k = 1 \mid X) = \sum_{l=1}^{L} q_l a_{lk} \tag{3.6}$$

where q_l = posterior probability for l = 1, …, L and

a_{lk} = attribute mastery pattern matrix, where k is the number of attributes.

Given that the estimated theta values, zeta values, and mahalanobis distances, the attribute mastery probabilities are calculated by using Equation 3.5 and 3.6. The RSM results are driven from post-hoc approach to provide diagnostic information for examinees in a mathematical test. This implies that RSM is not an approach of a cognitive model-based test design to incorporate cognitive theories.

3.5 Implementation of RSM

Tatsuoka (1991) has described an algorithm that generates all possible knowledge states from a given Q-matrix by using BUGLIB (Varadi, C. Tatsuoka, & K. Tatsuoka, 1992; Varadi & K. Tatsuoka, 1989) program. However, the computer program BUGLIB is limited to few researchers and not currently used for anyone who wants to apply or study RSM. Although the RSM is a well-known CDM and theoretically well-defined model, the lack of broadly used software is a crucial limitation of RSM. Simple algorithm with computer software should be developed to implement RSM.

CHAPTER 4

COGNITIVE DIAGNOSIS MODELS

In this chapter, CDMs were reviewed. They are Unified model, Fusion model, DINA, NIDA, DINO, NIDO, GDM, LLTM, MLTM, GLTM, LLTM-D, LCDM, AHM and BIN.

4.1 Unified Model

RSM is a pioneering method among the latent class cognitive models, but it has a limitation in figuring out sources of variation in response behavior from that predicted by Q. With an effort to produce a practical method for modeling and measuring cognitive aspects of examinees, the unified model also incorporates four fundamental sources of such response variation. The first one is Strategy Selection. According to Tatsuoka's Q-matrix (Tatsuoka, 1983), it is assumed that there is a single strategy for correctly answering each item, not multiple strategies. However, in reality, an examinee may or may not use the strategy based on the Q-matrix. Thus, the unified model considers this possibility and includes in the model as d_j which is the probability of selecting the Q-based strategy over all other strategies to solve an item. Second, Completeness is based

on an idea that an item *j* may require skills or attributes left out of the Q-matrix. In this case, the Q-matrix is said to be 'incomplete' for the item. Third source is Positivity. An examinee who "possesses" an attribute may not apply it successfully to a particular item that requires it. In contrast, an examinee who lacks the attribute may apply it correctly. When such non-Q-predicted behaviors are prevalent in the population, it can be said that the attribute is low positive for that item (DiBello, Stout, & Roussos, 1993). In the unified model, π_{ij} is defined as the probability of an examinee successfully applies attribute *i* to item *j* given that $a_j = 1$ for that examinee, and r_{ij} is defined as the probability of an examinee successfully applies attribute *i* to item *j* given that $a_j = 0$ that examinee. $\pi(1-r)$ is used as an index of positivity for the combination of attribute *i* and item *j* . As fourth source, slips indicate random errors committed by examines. Typos or lapses in attention can be examples of slips. This term differs from Tatsuoka's use in that any non-Q-predicted responses are regarded as slips in RSM, no matter what its cause (DiBello et al, 1993).

The item response function of the unified model is given in Equation (4.1).

$$P(X_i = 1 | \underline{\alpha}_j, \eta_j) = d_i \prod_{k=1}^{K} \pi_{ik}^{\alpha_{jk} \cdot q_{ik}} r_{ik}^{(1-\alpha_{jk}) \cdot q_{ik}} P_{c_i}(\eta_j) + (1 - d_i) P_{b_i}(\eta_j) \quad , \qquad (4.1)$$

where $P(X_i = 1 | \underline{\alpha}_j, \eta_j)$ is the probability of answering item *i* correctly given that examinee *j* has a skill mastery vector of $\underline{\alpha}_j$, and a supplemental ability parameter of η_j. α_j is obtained by calculating the proportion of the population that has mastered each skill *k*. The inclusion of the supplemental ability η_j implies that the Q-matrix is now necessarily a complete representation of all skill requirements of every item on the test

(Roussos, L. A., DiBello, L. V., Stout W., Hartz, S. M., Henson, R. A., & Templin, J. L., 2007). q_{ik} indicates whether or not skill *k* is required by item **i** on the Q-matrix. $P_h(\eta_j)$ is a Rasch model (Rasch, 1961) with the difficulty parameter equal to the negative of *h*, which can be expressed as.

$$P_h(\eta_j) = \{1 + \exp[-1.7(\eta_j + h)]\}^{1/2} \tag{4.2}$$

where *h* stands for either c_i in the first term or b_i is the second term. For each item *i* on the test, there are $2k_i + 3$ item parameters: d_i, π_{ik}, r_{ik}, two IRT Rasch Model parameters c_i and b_i. The unified model IRF not only take into account modeling examinee responses influenced by Q-matrix, but also reflect influence of examinee responses probability based on non-Q skills with the term $P_{c_i}(\eta_j)$ and alternative Q-strategies with the term $P_{b_i}(\eta_j)$ (Roussos, L. A., et al., 2007).

4.2 Fusion Model (Reparameterized Unified Model)

Although the unified model has cognitively interpretable parameters, unfortunately not all parameters are identifiable and thus, statistically estimable. For this reason, making a reparameterized version of the original fusion model was required by reducing the complexity of the parameter space. Hartz (2002) reduced the number of parameters from $2k + 3$ to $k + 2$ (k = number of skills required to solve an item). Therefore, the reduced model, referred to as the Reparameterized Unified Model (RUM), has a simple structure of the parameter space and enhanced ability to estimate the parameters compared to the original model. However, the RUM still maintains the

unified model's advantages of flexible capacity to fit diagnostic test data and retains most important components, such as the supplemental ability parameter η.

The Fusion model is mathematically equivalent to the original unified model, but strategy selection parameter (d_i) in the original unified model was omitted in the RUM by setting d_i parameter to 1. This means that no examinees may select other strategies than the Q-strategy to solve the item. Equation (4.3) presents the RUM IRF.

$$P(X_i = 1 | \underline{\alpha}_j, \eta_j) = \prod_{k=1}^{K} \pi_{ik}^{\alpha_{jk} \cdot q_{ik}} r_{ik}^{(1-\alpha_{jk}) \cdot q_{ik}} P_{c_i}(\eta_j) = \pi_i^* \prod_{k=1}^{K} r_{ik}^{*(1-\alpha_{jk}) \cdot q_{ik}} P_{c_i}(\eta_j), \quad (4.3)$$

$\pi_i^* (= \prod_{k=1}^{K} \pi_{ik}^{\alpha_{jk}})$ is the probability that an examinee having mastered all the Q required skills for solving item *i* will correctly apply all the skills to answer the item under the assumption of conditional independence of individual skill application. r_{ik}^* is a similar parameter to the positivity index in the unified model, but expressed as $\frac{p(Y_{ijk} = 1 | \alpha_{jk} = 0)}{p(Y_{ijk} = 1 | \alpha_{jk} = 1)} = \frac{r_{ik}}{\pi_{ik}}$ where π_{ik} is the probability that an examinee successfully applies attribute *k* to item *i* given that the examinee has mastered the attribute, and r_{ik} is the probability that an examinee successfully applies attribute *k* to item *i* given that the examinee has not mastered the attribute. The more strongly the item depends on mastery of a skill k_0, the closer $r_{ik_0}^*$ is to zero. Therefore, r_{ik}^* is like an inverse indicator of evaluating diagnostic capacity provided by item *i* about mastery skill *k* and reflect the influence of a skill on each individual item response probability (Roussos et al., 2007). The $P_{c_i}(\eta_j)$ component was retained from the unified model to consider that the Q-matrix

may not contain all relevant cognitive skills for solving an item. The $P_{c_i}(\eta_j)$ refers to the Rasch model with difficulty parameter $-c_i$ as shown in equation (3.2). c_i indicates the reliance of the missing multiple sills on the whole item response function. In the RUM, the lower the value of c_i, the lower the value of $P_{c_i}(\eta_j)$. If $P_{c_i}(\eta_j)$ is close to 0 for most examinees, the Q-matrix is incomplete and should be fixed. Thus, c_i provide important diagnostic information to check if a skill is missing from the Q-matrix or if the Q-matrix needs to be added more skills.

A Bayesian inference technique with the Markov chain Monte Carlo (MCMC) procedure has been adapted to estimating the item parameters (π_i^*, r_{ik}^*, c_i) and examinee skills parameters ($\underline{\alpha}_j$). MCMC algorithms are applicable to parametrically complex models such as the RUM than EM (expectation and maximization) algorithms. Moreover, one can obtain a joint estimated posterior distribution of both the test's item parameters and the examinee skill parameters from MCMC (Patz & Junker, 1999). The WINBUGS program (Spiegelhalter, Thomas, Best, & Lunn, 2003) and the Arpeggio program (Hartz, Roussos, & Stout, 2002) which is free can be used for parameter estimation of RUM. However, the MCMC analysis has heavy computational demand and uncertainty about the analysis result, especially with complex models requiring more parameters to estimate (Kim & Bolt, 2007).

In summary, the RUM is an attractive model to users because it has estimable parameters, but retains skills based interpretability of the unified model. Furthermore, the RUM provides useful information regarding item properties, Q-matrix, and examinees'

skill profiles. However, the RUM has a drawback due to the computational complexity of parameter estimation with MCMC.

4.3 DINA, HO-DINA, NIDA, DINO, and NIDO

The Deterministic Inputs, Noisy "And" gate (DINA; Haertel, 1989) model and the Noisy Input, Deterministic "And" gate (NIDA; Junker & Sijtsma, 2001) models are conjunctive (non-compensatory) models for skills diagnosis. On the other hand, the Deterministic Input; Noisy "Or" gate (DINO; Templin & Henson, 2006) model and the Noisy inputs, deterministic "or" gate (NIDO; Templin, 2006) are disjunctive models. Those four models (DINA, NIDA, DINO, & NIDO) are very similar in their item response functions.

4.3.1 DINA Model

In the DINA model, a latent variable ξ_{ij} for examinee *j* and *i*th item is defined as follows.

$$\xi_{ij} = \prod_{k=1}^{k} \alpha_{jk}^{q_{ik}} \tag{4.4}$$

If an attribute *k* is measured by an item *i*, then q_{ik} = 1. If an attribute is not measured, then q_{ik} =0. The term a_{jk} represents whether respondents have mastered the measured attribute. For example, suppose that there are four skills in a Q-matrix and skill 1 and 4 are required to solve an item. Then, for examinee *j*, $\xi_{ij} = a_{j1}^{1} \times a_{j2}^{0} \times a_{j3}^{0} \times a_{j4}^{1} = a_{j1} \times a_{j4}$. Because of the multiplicative term, the ξ_{ij} is equal to 1 only when both skills are present, which is the conjunctive aspect of the model. The vector of $\xi_{ij} = (\xi_{1j}, \xi_{2j}, ..., \xi_{ij})$ is the

deterministic input of the DINA model and the same concept as ideal response patterns in the rule space model.

However, there is possibility that respondents who have mastered all measure attributes (ξ_{ij}=1) incorrectly answer the item or respondents who have not mastered at least one of the measured attributes (ξ_{ij}=0) correctly answer the item. The former case is called "slip", and the later one is called "guess". The slipping and guessing parameter are expressed as follows:

$$g_i = P(Y_{ij} = 1 \mid \xi_{ij} = 0) \tag{4.5}$$

$$s_i = P(Y_{ij} = 0 \mid \xi_{ij} = 1) \tag{4.6}$$

where Y_{ij} is a response for examinee *j* and item *i*. From equations 5 and 6, the probability that an examinee gets an item correct is

$$P(Y_{ij} = 1 \mid \xi_{ij}) = (1 - s_i)^{\xi_{ij}} g_i^{1-\xi_{ij}} . \tag{4.7}$$

$(1 - s_j)$ is the probability of not slipping for item *j*. If an examinee has mastered all necessary attributes of an item, then the response probability is $(1 - s_j)^1 g_j^{\ 0} = (1 - s_j)$. If this person has not mastered the attributes, $P = (1 - s_j)^0 g_j^{\ 1} = g_j$. Because X_{ij} is considered a Noisy observation of each ξ_{ij} and the binary value of ξ_{ij} (1 or 0) influence on selecting between the probabilities $(1 - s_j)$ and g_j, this model is called Deterministic Inputs, Noisy "And" gate model.

4.3.2 Higher-Order DINA Model

Higher-order DINA model (HO-DINA; de la Torre & Douglas, 2004) assumes that attributes are hierarchically structured. Hierarchy relationships of attributes differ from pre-requisite relationships of attributes. A hierarchy relationships means that attributes are ordered in difficulty. In the higher-order formulation, the components of attribute vector ($\bar{a}_j$) is assumed to be conditionally independent given a higher-order latent proficiencyθ_j (de la Torre & Lee, 2010). The probability model for $\bar{a}_j$ conditional on θ_j is

$$\mathrm{P}(\bar{a}_j|\theta_j) = \prod_{k=1}^{K} \mathrm{P}(a_{jk}|\theta_j). \tag{4.8}$$

The probability of mastery can be expressed using the following latent logistic regression model:

$$\mathrm{P}(a_{jk} = 1|\theta_j) = \left\{\frac{\exp[1.7\lambda_1(\theta_j - \lambda_{ok})]}{1+\exp[1.7\lambda_1(\theta_j - \lambda_{ok})]}\right\} \tag{4.9}$$

where $\lambda_1$and λ_{ok} represent the latent discrimination and difficulty parameters, respectively. The constant 1.7 was added to give the λ_s similar interpretations as the difficulty and discrimination parameters in IRT models (Torre & Douglas, 2008). In this formulation, attributes with higher λ_{ok} are regarded as more difficult to master.

4.3.3 NIDA Model

Even though aberrant responses are modeled in the DINA model, it does not differentiate between respondents who lack only one of the measured attributes and those who not mastered any of the attributes. In contrast to the DINA model, the noisy-input, deterministic-and-gate (NIDA) model (e.g., Junker & Sijtsma, 2001) has one slipping parameter and one guessing parameter per attribute, not per item.

$$g_k = P(\xi_{ijk} = 1 \mid a_{jk} = 0) \tag{4.8}$$

$$s_k = P(\xi_{ijk} = 0 \mid a_{jk} = 1) \tag{4.9}$$

The value of ξ_{ijk} indicate whether respondent *j* correctly apply attribute *k* for item *i* ($\xi_{ijk} = 1$) or not ($\xi_{ijk} = 0$). The parameter a_{jk} is an indicator of attribute mastery for examinee *j* as in the DINA. The guessing parameter (g_k) is the probability of the correct application of attribute *k* in the context of item *i* even though the attribute has not been mastered. Likewise, the slipping parameter (s_k) is the probability of the incorrect application of attribute *k* in the context of item *i* even though the attribute has been mastered. These parameters are defined at the level of attributes, thus they increase with the number of attributes. The final item response function of the NIDA model is

$$P(Y_{ij} = 1 \mid \underline{\bar{a}}, s, g) = \prod_{k=1}^{K} P(\xi_{ijk} = 1 \mid a_{jk}) = \prod_{k=1}^{K} [(1 - s_k)^{a_{jk}} g_k^{1-a_{jk}}]^{q_{ik}} \tag{4.10}$$

where X_{ij} is the observed response for item *j* and examinee *i*, q_{jk} indicates whether attribute *k* is measure by item *i* in the Q-matrix, and $(1 - s_k)$ is the probability of not slipping for attribute *k*.

4.3.4 DINO Model

The deterministic input, noisy-or-gate (DINO) model (Templin & Henson, 2006) is the compensatory version of the DINA model. While the DINA is a conjunctive ("And" gate) model, the DINO is a compensatory ("Or" gate) model. As in the DINA model, there is a gate component in the DINO model. The latent response variable (ω_{ij}) is defined for examinee *j* and item *i*.

$$w_{ij} = 1 - \prod_{k=1}^{k} (1 - a_{jk})^{q_{ik}} \quad (4.11)$$

where a_{jk} and q_{ik} are the same indicators as in the DINA. If an attribute *k* is measured by an item *j* and an examinee *i* possess the attribute, $(1 - a_{jk})^{q_{ik}} = (1-1)^1 = 0$. If an examinee does not possess the measured attribute, $(1 - a_{jk})^{q_{ik}} = (1-0)^1 = 1$. Because the product term is defined over all attributes, w_{ij} is 1 if the examinee has satisfied at least one attribute (e.g., symptom) for the item. Only when all attributes in the Q-matrix are not present, w_{ij} is 0. The DINO model also has the slipping and guessing parameters as in the DINA as follows.

$$g_i = P(Y_{ij} = 1 \mid w_{ij} = 0) \quad (4.12)$$

$$s_i = P(Y_{ij} = 0 \mid w_{ij} = 1) \quad (4.13)$$

The probability of correct response for examinee *j* and item *i* is defined as:

$$P(Y_{ij} = 1 \mid \omega_{ij}) = (1 - s_i)^{\omega_{ij}} g_i^{1-\omega_{ij}} . \quad (4.14)$$

If all measured attributes are absent, the probability is $(1 - s_i)^0 g_i^{1-0_{ij}} = g_i$, and if at least one attribute is present for the examinee, $P(Y_{ij} = 1 \mid \omega_{ij} = 1) = (1 - s_i)^1 g_i^{1-1} = (1 - s_i)$. Since the DINO is a compensatory model, it does not matter how many or which particular attributes possess for an examinee (DiBello et al., 2007).

4.3.5 NIDO Model

The noisy input, deterministic-or-gate (NIDO) model is the compensatory version of the NIDA model. The NIDO model estimates one intercept and one slope parameter for each attribute but with equality constraints across items (Rupp, et. al., 2010). The

concept of slope and intercept in the NIDO is originated from regression analysis. In order to build the NIDO model, first of all, a "kernel" should be expressed with the intercept and slope parameters as follows:

$$\ker nel_j = \sum_{k=1}^{k} (\lambda_{.,0,(k)} + \lambda_{.,1,(k)} \alpha_{jk}) q_{ik} \tag{4.15}$$

Where the term α_{jk} is an indicator of whether respondents have mastered the measured attribute or not for examinee *j*, and q_{ik} represents involvement of attribute *k* in the Q-matrix for item *j*. The first subscript in the intercept ($\lambda_{.,0,(k)}$) and slope ($\lambda_{.,1,(k)}$) parameter represent the item to which the parameter corresponds (Rupp, et. al., 2010). It is simply expressed as a "dot" because all parameters for attributes are equal across items. The second subscript determines the characteristic of the parameter; 0 for intercept and 1 for slope. The denotation of 0 or 1 is similar to the dummy-coded independent variables in the dummy-coded analysis of variance. Thus, the slope parameter $\lambda_{.,1,(k)}$ can be main effects in the NIDO model. With the kernel, the formula for the NIDO model is shown as

$$P(Y_{ij} = 1 | \overline{a_{jk}}) = \frac{\exp(\sum_{k=1}^{k} (\lambda_{.,0,(k)} + \lambda_{.,1,(k)} \alpha_{jk}) q_{ik})}{1 + \exp(\sum_{k=1}^{k} (\lambda_{.,0,(k)} + \lambda_{.,1,(k)} \alpha_{jk}) q_{ik})} \tag{4.16}$$

where Y_{ij} is the observed response for examinee *j* and item *i*, *P* is the probability of correct response, exp( ) is the exponential function (approximately 2.718 raised to the power of the terms in the parentheses), the term in the parentheses is the kernel described above.

4.4 GDM

Among the family of cognitive diagnostic models, the rule space model (Tatsuoka, 1983) is an approach based on IRT model. Some other CDMs (Haertel, 1989; Maris, 1999) were developed by extending latent class analysis. The HYBRID model (Yamamoto, 1989), the mixed Rasch model (Rost, 1990; von Davier & Rost, 1995) and mixture IRT models (Keldermann & Macready, 1990; Mislevy & Verhelst, 1990) integrate latent class approaches and IRT. The class of general diagnostic models (GDMs; von Davier, 2005) is also an extension of IRT and latent class models.

However, GDMs extend the applicability of skill profile models to confirmatory multidimensional models with discrete latent trait variables. Polytomous items and skills with more than two proficiency levels can be dealt with in the GDMs framework. Furthermore, both compensatory and non-compensatory models can be specified within the GDMs. Multivariate versions of the Rash's (1960) model, the two-parameter logistic item response theory (2PL IRT) model (Birnbaum, 1968), the generalized partial credit model (GPCM; Muraki, 1992) are examples of the GDMs for partial credit data. Therefore, GDMs can be said as integrated version of the areas of IRT, latent-class analysis (Lazarsfeld & Henry, 1968) and multiple classification latent-class models (Goodman, 1974).

Like many of the other CDMs, the class of GDMs also uses Q-matrix, but entries in the Q-matrix can include polytomous item responses and polytomous attributes. In the GDMs, specification of Q-matrix allows researchers to see how skill patterns and the Q-matrix interact. The GDM allows ordinal skill levels and different forms of skill dependencies to be specified (von Davier, 2005a). Therefore, more gradual differences between examinees can be modeled in this framework (von Davier, 2010).

The GDM can be specified with a logistic model (Von Davier, 2005a). The class of GDMs are defined for dichotomous and polytomous data and with the ability to model multiple, potentially mixed, dichotomous and ordinal skill variables. Assume N examinees are sampled with observations on *I* discrete response variables $\underline{x_n} = (x_{n1}, x_{n2}, ..., x_{nI})$, each with outcomes $\underline{x_n} \in \{0, 1, \ldots, \mathrm{m_i}\}$. Assume there is a set of additional K discrete random variables $\underline{a_n} = (a_{n1}, a_{n2}, ..., a_{nI})$, with realizations $\underline{a_{nk}} \in \{\mathrm{s_k}(0), \ldots, \mathrm{s_k}(\mathrm{l_k})\}$. which are unobserved for each examinee $n = 1, ..., N.$: The $\mathrm{s_k}(\mathrm{l})$ are skill levels of the *kth* unobserved skill variable. Typical choices of skill levels as they are common in IRT models and profile scoring models are discussed in the subsequent section. The $\underline{a_n}$ constitute the multidimensional latent variable outcome (referred to as, for example, skill profile or attribute pattern, or multiple proficiencies in the literature) for examinee n, and are the target of inference in the following exposition. Assume that the $\overline{x_n}$ are independent and identically distributed with

$$P(x_{n1}, ..., x_{nI}) = \sum_g P(g) \int_{a=(a_1,...,a_k)} p(\underline{a} \mid g)\, p(a_{n1}, ..., a_{nI} \mid \underline{a}, g) d\underline{a}, \qquad (4.17)$$

where $P(\underline{a} \mid g)$ denotes the distribution of the unobserved variable in population g and $P(x_{n1}, ..., x_{nI} \mid \underline{a}, g)$ denotes the conditional distribution of the vector of observed variables $(x_{n1}, x_{n2}, ..., x_{nI})$ given unobserved variable $(a_1, a_2, ..., a_k)$ and population *g*. Equation (4.17) presents the marginal probability of a vector of observed variables based on the conditional distribution given skill variables and allows for more than one population. The populations are defined by either manifest grouping variables or as in mixture distribution IRT models (Mislevy & Verhelst, 1990; Rost, 1990; von Davier & Rost,

1995; Yamamoto, 1989), or a combination of observed and unobserved data in the mixing/population indicator (von Davier & Yamamoto, 2004c).

For the general diagnostic model, local independence of the components x_{nI} given $\underline{a}$ is assumed, which yields

$$P(x_{n1},...,x_{nI} \mid \underline{a}) = \prod_{i=1}^{I} P_i(x_{ni} \mid \underline{a}), \tag{4.18}$$

so that the conditional probability of a vector of observed responses can be written as a product of conditional response probabilities of each of the components of the response vector.

Let Q is a binary I x K matrix where I represents observed response variables, and K represents skill variables. That is $Q = (q_{ik})$ where $i = 1,...,I, J = 1,...,K,$ and $q_{ik} \in \{0,1\}$. Like in RSM, Q-matrices are used to determine ideal patterns of observed responses given a specific skill profile $\underline{a} = (a_1, a_2,..., a_k)$. Probabilistic models for cognitive diagnosis use this matrix to specify the conditional probability of the observed response vector given the latent variable $\underline{a}$. Finally, the class of GDM is given by

$$P_i(x \mid \underline{a}) = P(x \mid \beta_{xig}, q_i, \gamma_{ig}, \underline{a}) = \frac{\exp[\beta_{xig} + \gamma_{xig}^T h(q_i, \underline{a})]}{1 + \sum_{y=1}^{m_i} \exp[\beta_{xig} + \gamma_{xig}^T h(q_i, \underline{a})]} \tag{4.19}$$

with real-valued difficulty parameters β_{xig} in population *g* and with a *k*-dimensional slope parameter $\gamma_{xig} = (\gamma_{xig1},...,\gamma_{xigK})$ (used in transposed form γ^T in the equation) for written as a multiple group or a discrete mixture model. Multiple group GDMs may be estimated by concurrent calibration with completely separate parameter sets or with parameter sets that include equality constraints or fixed parameters across groups.

As extensions of GDM, Von Davier also introduced the mixture general diagnostic models (MGDM; Von Davier, 2008b) which is the discrete mixture distribution version of the GDM and the hierarchical general diagnostic model (HGDM; von Davier, 2007). The mdltm softwere (Von Davier, 2005b) is used to estimate MGDMs and HGDMs.

4.5 LLTM

Latent class models are appropriate to obtain classification results (e.g., attribute mastery probabilities) for examinees' skill diagnosis. Unfortunately, those models cannot be used for test design based on cognitive theory except AHM. However, most latent trait models can test the significance of item parameter estimates in the model, and thus, it is possible to construct a test with items that reflect a cognitive theory. Q-matrix is also used for latent trait models. The linear logistic test model (LLTM; Fischer, 1973) was the first psychometric model that links cognitive psychology to item design.

LLTM is a generalization of the Rasch model and includes item stimulus features as cognitive variables to predict item success. The Rasch model is expressed as

$$P(X_{is}=1)=\frac{\exp(\theta_s-\beta_i)}{1+\exp(\theta_s-\beta_i)}, \tag{4.20}$$

where $P(X_{is}=1)$ is the probability that person *s* passes item *i*, θ_s is the trait level of person *s*, and β_i is the difficulty parameter of item *i*. Unlike the Rasch model, in LLTM β_i is replaced with a linear function to include cognitive variables:

$$\beta_i=\sum_{k=0}^{K}\eta_k q_{ik}=\eta_0 q_{i0}+\eta_1 q_{i1}+\eta_2 q_{i2}+\ldots+\eta_k q_{ik}, \tag{4.21}$$

where η_k represents the effect of stimulus feature *k*, q_{ik} is the score (e.g., 0 = absence and 1 = presence) of stimulus feature *k* of item *i*, and $\eta_0 q_0$ is the intercept of the equation. Finally, the full LLTM is expressed as;

$$P(X_{is}=1)=\frac{\exp(\theta_s-\sum_{k=0}^{K}\eta_k q_{ik})}{1+\exp(\theta_s-\sum_{k=0}^{K}\eta_k q_{ik})}. \quad (4.22)$$

The ***Q*** matrix which consists of q_{ik} has a *I* x *K* matrix (*K* < *I*). The elements of q_{ik} are dichotomous scores for the item on cognitive model variables.

4.6 MLTM, GLTM, & MLTM-D

MLTM (Whitely, 1980), GLTM (Embretson, 1984), and MLTM-D (Embretson, 2013) are non-compensatory multidimensional IRT models.

4.6.1. MLTM

Most achievement test items require multiple skills or competencies to obtain correct responses. The multidimensional latent trait model (MLTM; Whitely, 1980; Embretson, 1991; Embretson & Yang, 2006) combines a mathematical model of item accuracy with individual differences model. MLTM is based on a continued product of processing outcome probabilities to accommodate any number of subtasks. The subtasks are also called components in MLTM. MLTM estimates multiple components difficulties (item parameters) and multiple component trait levels (person parameters). MLTM is given as follows:

$$P(X_{is}=1)=\prod_k P(X_{isk})=\prod_k \frac{\exp(\theta_{sk}-\beta_{ik})}{1+\exp(\theta_{sk}-\beta_{ik})}, \quad (4.23)$$

where $P(X_{is}=1)$ is the probability that person *s* performs successfully on item *i* and $\prod_k P(X_{isk})$ is the product of the probabilities that the examinee success on each processing component *k*, given the correct outcome of the preceding component. The right side of the equation contains Rasch models for the probability of success on each component, where θ_{sk} is the trait level of person *s* on component *k* and β_{ik} is the difficulty of item *i* on component *k*.

4.6.2. GLTM

The general component latent trait model (GLTM; Embretson, 1984) is a generalized form of this model in which β_{ik} term is replaced by the following equation.

$$\beta_{ik} = \sum_{m=0}^{m} \eta_{km} q_{ikm} \tag{4.24}$$

where q_{ikm} is the score of stimulus feature *m* on component *k* for item *i*, η_{ikm} is the weight of stimulus feature *m* on component *k*, and $\eta_{k0} q_{ik0}$ is an intercept. That is, β_{ik} indicates the weighted sum of underlying stimulus factors to represent scored attributes. The full GLTM is followed as:

$$P(X_{isT}=1) = \prod_k \left[\frac{\exp(\theta_{sk} - \sum_{m=0}^{m} \eta_{km} q_{ikm})}{1+\exp(\theta_{sk} - \sum_{m=0}^{m} \eta_{km} q_{ikm})} \right]. \tag{4.25}$$

The GLTM enables an examination of how the underlying stimulus features will impact the difficulty of each component (β_{ik}) based on pre-established cognitive theories. Since GLTM is an extension of the MLTM, it also estimates individual ability on each component (also called cognitive attribute) as a continuous variable, thus giving detailed

information about an examinee's skill profile. Like MLTM, GLTM requires two kinds of data, responses to total items and responses to a series of each component.

4.6.3. MLTM-D

MLTM-D is a diagnostic, noncompensatory IRT model which has two levels of a hierarchical structure, components and attributes. This model is appropriate when two or more components affect problem solving and the difficulties of these components are influenced by those of attributes. The model identification of MLTM-D is based on the structure of the component matrix (**C**) and the attribute matrix (**Q**) containing independent vectors of item scores. These matrices are constructed based on a plausible theory. That is, the component and attribute matrices can be defined only when a theory is available. The **C** and **Q** matrix show the relationships between the components and specific items as well as the relationships between the attributes and the components. In these matrices, if an item is related with two components among all the identified components, the corresponding two elements in the matrix indicate 1; otherwise 0. The attributes are indicated for each component in the same way. The probability of solving item *i* successfully by examinee *j*, $P(Y_{ij}=1)$, in MLTM-D is shown as follows:

$$P_{ij} = P(Y_{ij}=1) = \prod_{m=1}^{M} P_{ijm}{}^{C_{im}} \tag{4.26}$$

and $$P_{ijm} = P(Y_{ij}=1 \mid \theta_{jm}, \underline{q}_{im}, \underline{\eta}_{m}) = \frac{\exp\{-1.7(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0})\}}{1+\exp\{-1.7(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0})\}}, \tag{4.27}$$

where θ_{jm} indicate the ability of person j on component *m*, q_{imk} is the score of attribute *k* in components *m* for item *i*, η_{mk} is the weight of attribute *k* on component *m*, η_{mo} is the intercept for component *m*, and c_{im} is the involvement of component *m* in item *i*.

The ability estimates are obtained for each component, and the level of the component is determined by the specific attributes relevant to the component. MLTM-D provides two levels of diagnostic information: the trait level estimates on each component dimension, θ_{jm}, and skill mastery of attributes underlying the difficulty of each component, q_{imk}. SAS and SPSS macros can be used for the person or item parameter estimation with MLTM-D (Embretson, 2013).

4.7 LCDM

The log-linear cognitive diagnostic model (LCDM; Henson et al., 2009) is a generalized model to express both conjunctive and disjunctive models. In LCDM, a non-compensatory model is viewed as a model where the relationship between any attribute required in an item and the item response (*x*) depends on mastery or nonmastery of the remaining other required attribute. On the other hand, a compensatory model is a model where there is no conditional relationship between the attributes and *x* (Henson et al., 2009). The general form of the LCDM is as follows:

$$P(X_{ri} = 1|\boldsymbol{\alpha_r} = \boldsymbol{\alpha_c}) = \frac{\exp[\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_r)]}{1+\exp[\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_r)]}, \quad (4.27)$$

where $P(X_{ri} = 1|\boldsymbol{\alpha_r} = \boldsymbol{\alpha_c})$ is the probability that respondent *r* with the attribute-mastery profile $\boldsymbol{\alpha_c}$ correctly responds to the i^{th} item. $\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_r)$ can be rewritten as:

$$\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_r) = \lambda_{i,0} + \sum_{u=1}^{K} \lambda_{i,1,(u)}(\alpha_{ru} q_{iu})$$

$$+\sum_{u=1}^{K}\sum_{v>u}\lambda_{i,2,(u,v)}(\alpha_{ru}\alpha_{rv}q_{iu}q_{iv})+\cdots, \quad (4.28)$$

where, $\boldsymbol{\lambda}_i^T$ is a $1 \times (2^K - 1)$ vector of weights (k = # of attributes) for the i^{th} item. For example, $\lambda_{i,1,(1)}$ represents a simple main effect of attribute 1, $\lambda_{i,1,(2)}$ refers to a simple main effect of attribute 2, and $\lambda_{i,2,(1,2)}$ represents a two-way interaction of attributes 1 and 2. $\lambda_{i,0}$ is an intercept. $\boldsymbol{q}_i$ is the Q-matrix entries of attributes to be measured in the i^{th} item ($k \times 1$ vector). $\boldsymbol{\alpha}_r$ represents the attribute mastery profile of respondent r ($1\times k$ vector). $\boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_r)$ is a set of linear combinations of $\boldsymbol{q}_i$ and $\boldsymbol{\alpha}_r$. Therefore, the probability of a correct response for an item which requires two attributes (A_1 and A_2) can be defined as:

$$P(X_{ri}=1|\boldsymbol{\alpha_c})=\frac{\exp[\lambda_{i,0}+\lambda_{i,1,(1)}(\alpha_1)+\lambda_{i,1,(2)}(\alpha_2)+\lambda_{i,2,(1,2)}(\alpha_1\alpha_2)]}{1+\exp[\lambda_{i,0}+\lambda_{i,1,(1)}(\alpha_1)+\lambda_{i,1,(2)}(\alpha_2)+\lambda_{i,2,(1,2)}(\alpha_1\alpha_2)]}. \quad (4.29)$$

In this equation, if attribute 1 (A_1) is mastered (α_1=1), then the probability of a correct response increases by a factor of $e^{\lambda_{i,1,(1)}}$ given that other attribute (attribute 2) has not been mastered. $\lambda_{i,2,(1,2)}$ represents the extent to which the conditional relationship of A_1 and the item response depends on attribute 2 (A_2). Thus, if A_2 is mastered (α_2=1), the probability of a correct response increases by a factor of $e^{\lambda_{i,1,(2)}+\lambda_{i,2,(1,2)}}$. Such a model in Equation 3.28 can be extended to include all possible main effects and interactions of attributes.

One important advantage of LCDM is that this model can provide empirical information regarding the relationship between attribute mastery and the item response patterns without specification of a type of model such as compensatory or non-compensatory (Henson et al., 2009). Therefore, what type of model could have better fit for some test items can be confirmed with LCDM. However, disadvantage of LCDM is that the number of item parameters is multiplied by two as one attribute is added to the model. In other words, 2^k item parameters are needed when the model has k attributes.

Such large number of item parameters (2^k) causes even heavy computational demand and have large standard errors for the parameter estimation.

4.8 AHM

The Attribute Hierarchy Method (AHM; Gierl, 2007; Gierl, Leighton, & Hunka, 2007; Gierl, Alves, & Majeau, 2010; Leighton, Gierl, & Hunka, 2004) is a framework for designing diagnostic items based on attributes which have a hierarchical structure. AHM is similar to RSM in that both models use attributes and Q-matrix as a cognitive model. Furthermore, expected response patterns which are hypothetically generated are used to classify examinees' response patters in AHM as well as RSM. However, AHM differs from RSM with the assumption of dependencies among the attributes within the cognitive model. Modeling cognitive attributes using AHM necessitates the specification of a hierarchy outlining the dependencies among the attributes. In contrast, RSM makes no assumptions regarding the dependencies among the attributes. This difference has lead to the development of both IRT and non-IRT based psychometric procedures for analyzing test item responses using AHM. AHM also differs from RSM with respect to the identification of the cognitive attributes and the logic underlying the diagnostic inferences made from the statistical analysis (Gierl, 2007). That is, RSM uses a post-hoc approach to the identification of the attributes required to successfully solve each item on an existing test, while AHM uses an a priori approach to identifying the attributes and specifying their interrelationships in a cognitive model.

AHM can be described with four step process. First of all, an AHM analysis begins with the specification of a cognitive model of task performance. Then, these

attributes are structured using a hierarchy so the ordering of the cognitive skills is specified. AHM assumes that test performance depends on a set of hierarchically ordered attributes (Gierl, et. al., 2007). Likewise in RSM, examinees are supposed to possess the attributes to answer test items correctly.

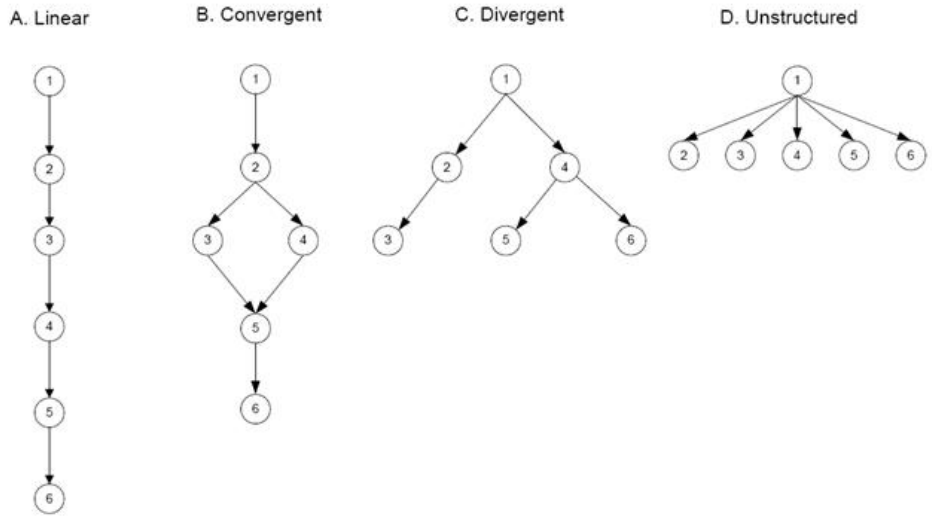

Figure 6. Visual Representation of the Four Different Hierarchical Structures

Leighton et al., (2004) identified four general forms of hierarchical structures in a cognitive model, as shown in Figure 6 (A) represents the linear hierarchy; attribute 1 is prerequisite to attribute 2, and attributes 1 and 2 are prerequisite to attribute 3.Attributes 1 to 5 are considered prerequisite to attribute 6. The attribute 6 is the most complex one because its existence depends on all other attributes in the model. (B) represents the convergent hierarchy; attributes 1 and 2 are prerequisite to attribute 3 and 4, but 3 or 4 are prerequisite to 5. (C) represents the divergent hierarchy which have multiple end states consisting of two or more attributes. (D) represents the unstructured hierarchy. Educational tests that clearly measure different knowledge components could be

characterized as an unstructured hierarchy of attributes because there is no ordering among the attributes (Gierl, et. al., 2007). These forms of hierarchical structures can be expanded and combined to form increasingly complex networks of hierarchies.

Next, a cognitive model of task performance should be constructed to make inferences about examinees' cognitive skills. The quality of diagnostic inferences produced with AHM depends on accuracy of the cognitive model of task performance (Gierl, Leighton, & Hunka, 2007). The cognitive model specified by hierarchy structures must be identified prior to developing a test because it is a guideline to develop test items (Gierl, Leighton, & Hunka, 2000). After identifying hierarchical relationship among the attributes, the Q-matrix is constructed by specifying all possible combinations of the attributes. Boolean additions are used to create the Q-matrix. The Q-matrix is used as a test blueprint to develop items that measure specific attributes outlined in the hierarchy structure (Gierl, et. al., 2007). Test items are developed by using the Q-matrix. Since this test specification is pre-determined, cognitive theory can have a clearly defined role in the test design. Misspecification of the Q-matrix or less impact of cognitive theory on test design can be avoided with this characteristic of the AHM.

In the third step, Given the Q-matrix and the attributes that examinees posses, the expected response patterns can be obtained. The hypothetical examinees are assumed to apply the attributes systematically based on the hierarchical structure. Real examinee's response patterns are classified by using a set of unique expected response patterns. If an observed response pattern of an examinee matches with an expected pattern, a "match" is noted and it can be inferred which attributes the examinee possesses. However, if the expected pattern is not logically included in the observed pattern, meaning that there are

unmatched patterns, the likelihood of misfits should be considered. The misfit indicates the case where the examinees possess all the attributes to solve an item, but answer incorrectly or the opposite case by guessing. These misfits can be occurred in many testing situations. However, if there are many discrepancies between the expected patterns and the observed patterns, then it should be considered that the attributes were accurately identified, the hierarchy was appropriately specified, the items measure the attributes, or the test was appropriate for the sample (Gierl, et. al., 2007).

The fit of the cognitive model relative to the observed response patterns can be evaluated using the Hierarchical Consistency Index (HCI; Cui, Leighton, Gierl, & Hunka, 2006). The HCI evaluates "the degree to which the observed response patterns are consistent with the attribute hierarchy" (Cui, et. al., 2006). The *HCI* for examinee *i* is given by:

$$HCI_i = 1 - \frac{\sum_{j=1}^{J} \sum_{g \in s_j} X_{i_j}(1 - X_{i_g})}{N_{c_i}} \qquad (4.8.1)$$

where J is the total number of items, X_{ij} is examinee i 's score (i.e., 1 or 0) to item j, S_j includes items that require the subset of attributes of item j, and N_{ci} is the total number of comparisons for correctly answered items by examinee i. The values of the HCI range from -1 to +1. Values closer to 1 indicate a good fit between the observed response pattern and the expected examinee response patterns generated from the hierarchy. Conversely, low HCI values indicate a large discrepancy between the observed examinee response patterns and the expected examinee response patterns generated from the hierarchy. HCI values above 0.70 indicate good model-data fit. If the data is not shown to fit the model, then various reasons may account for the discrepancies including: a misspecification of the attributes, incorrect ordering of attributes within the hierarchy,

items not measuring the specified attributes, or the model is not reflective of the cognitive processes used by a given sample of examinees. Therefore, the cognitive model should be correctly defined to make appropriate inferences for the examinees' knowledge and skills.

The last step is to report diagnostic results using AHM. Unfortunately, specific procedures for the cognitive score reporting are not clearly documented yet in AHM framework. Adopting classification procedures in RSM would be a way to solve this limitation of AHM. The details regarding classification in RSM will be described in the next section.

4.9 BIN

The idea of hierarchical factor structures in multidimensional FA models has been adapted in applications of Bayesian inference networks *(*BIN; Levy & Mislevy, 2004). In BIN, even the higher-order latent variables are typically categorical so that these models represent a completely categorical hierarchy of latent variables (Yan, Mislevy, & Almond, 1993). In the cognitive diagnostic framework, Bayesian networks shows hierarchically structured and graphical representation of probabilistic relationships between several attributes.

Bayesian networks are mechanisms for applying Bayes' theorem to complex cognitive problems. Bayes' theorem is based on the calculation of conditional probabilities to predict the chance of an event. Thus, bayseian inference networks (BINs) can be considered an entire class of statistical models with a full probability structure to CDMs (Rupp, et al., 2008). As a tool of Bayesian statistics, Bayes' theorem is easy to

understand and can be used for all situations. The mathematical statement of Bayes' theorem is expressed as

$$P(\theta_i|D) = \frac{P(D \cap \theta_i)}{P(D)} = \frac{P(\theta_i) \times P(D|\theta_i)}{\sum_i P(\theta_i) \times P(D|\theta_i)} \quad (4.9.1)$$

where θ_i is an unobservable event (parameter),

D is an observable event (data),

$P(\theta_i)$ is the prior probability of event θ_i,

$P(D|\theta_i)$ is the likelihood (conditional probability) of D given θ_i,

$\sum_i P(\theta_i) \times P(D|\theta_i)$ is the marginal probability of D, and

$P(\theta_i|D)$ is the posterior probability of θ_i given D.

The purpose of applying BINs to diagnostic assessment is to classify respondents by using multiple latent variables that represent the attributes to be measured (Rupp, et al., 2008). Under the Bayesian estimation approach, the classification results can be obtained by computing the posterior distribution for the parameters which combine the prior information. One important advantage of Bayesian approach over the classical approach to statistics is that it allows inference probabilities about the unknown parameters are based on the actual occurring data (Bolstad, 2007). Another advantage of BINs is that they can be used to specify complex attribute structures and the estimation of parameters in the model (Rupp, et al., 2008). However, the estimation for large sample sizes requires well-specified prior information.

CHAPTER 5

APPLICATION OF CDMS ACROSS COGNITIVE DOMAINS

5.1 Issues in the Applications of CDMs

Cognitive diagnostic assessment (CDA) is a new theoretical framework for psychological and educational testing that is designed to diagnose detailed information about examinees' mastery of attributes in a test (specific knowledge structures and processing skills). During the last three decades, many psychometric models have been developed for CDA. When applying those CDMs to a particular area of measurement settings, following issues should be considered.

5.1.1 The Definitional Grain Size of the Attributes

Attributes are essential components in the applications of CDMs because the design of the diagnostic assessment is based on the specification of the Q-matrix consisting of attributes. When decomposing cognitive constructs that try to measure in a test, the degree of definitional specificity is an important issue. This is often referred to as the definitional grain size of the attributes (Rupp, et al, 2009). Choosing a degree of the definitional grain size is a trade-off because each grain size has advantages and drawbacks. Finely defined attributes for tasks that are restricted in scope are required to make specific inferences regarding examinees' cognitive skills (Gierl & Leighton,

2007a). However, a cognitive analysis at a fine grain size may limit depth of construct representation and content coverage. On the other hand, coarsely defined attributes for tasks that are broader in scope may sacrifice specifications for lower levels of cognitive skills that are required in the assessments. In general, course-grained descriptions of attributes are used in broad blueprints for educational assessments while fine-grained behavioral and process descriptions are used in standards-based assessments that are geared to aligning the goals of curriculum, instruction, and assessment (Gierl & Leighton, 2007a). CDAs should be specified at the appropriate grain size, and its appropriateness depends on the objective of the CDA and type of reporting methods used in the assessment (Gierl & Leighton, 2009). Researchers need to determine a statistically manageable number of attributes by considering length of the tests and number of respondents. The more attributes are involved in a Q-matrix, the more number of parameters should be estimated in a given CDM. The large number of parameters causes heavy computational demands for the estimation because the number of parameters increases as a proportion of the number of attributes (= k) to be measured in an item. For example, fusion model has k+2 parameters, the DINO model includes 2k parameters, and RSM and LCDM has even 2^k parameters.

5.1.2 Reporting Cognitive Diagnostic Results

As an interface between test developer and users, score reporting conveys information regarding the meanings of test results. The test results are psychological outcomes from sophisticated assessments based on CDMs. It is an important issue for test developers to communicate with different types of test users with appropriate

interpretations of test outcomes. Standards claimed by National Council on Measurement in Education (1999) indicate that “the interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used”. Another important factor for successful score reporting is to minimize lag time between taking a test and reporting results without sacrificing the quality and integrity of the assessments (Huff & Goodman, 2007). Appropriately reported diagnostic results can be used in many different ways. Teachers can use the results to reflect on their instructional methods and to plan remedial activities to help individual students. Teachers can also use accumulated test results to communicate with parents, school administrators, and educational policy makers to enhance the quality of education and allocate the necessary resources properly. Students can know their strength and weakness on specific skills or knowledge through the test results, and in turn they can enhance their overall performance by making their effort on the area shown weak performance. Teachers can advise their students to change their learning strategies or use supplemental tutorials if necessary.

5.2 Software for CDMs

CDMs have a lot of advantages over other psychometric models. Although they have successfully provided useful diagnostic information about the examinees, utilizing in practice is complex. Unfortunately, the parameter estimation with CDMs requires complex statistical computations. For some CDMs (e.g., Fusion and LCDM), MCMC algorithm is used for the parameter estimation because it is easier to extend to

parametrically complex models than Expectation Maximization (EM) algorithms. However, the MCMC causes even heavier computational demand than the marginal maximum likelihood estimation (MMLE) with the EM algorithm. It takes several hours even for a single estimation and a day or more for more complex models or large amounts of data. Also, the MCMC can be misused easily because of the complexity of its algorithms. Furthermore, most software used for the applications of CDMs are not user-friendly. Table 3 shows currently available software for estimating CDMs. More user-friendly software or easy algorisms modified from previous methods should be developed in future research. Improvement of methodological tools of CDMs will lead to broad use of CDMs in the field of educational and psychological measurement.

Table 3 Software for Estimating CDMs

Software	*Type of software(Contact)*	*Estimated models*
Arpeggio	Commercial(www.assess.com)	Full NC-RUM, reduced NC-RUM
DCM	Freeware (requires the commercial version of Mplus)(jtemplin@uga.edu)	DINA, NIDA, DINO, NIDO, reduced NC-RUM, C-RUM
DCM in R	Freeware (requires freeware R) (alexander.robitzsch@iqb.hu-berlin.de)	DINA, DINO
DINA in Ox	Freeware (requires freeware Ox) (j.delatorre@rutgers.edu)	DINA
LCDM	Freeware (requires the commercial version of Mplus)(jtemplin@uga.edu)	LCDM family
MDLTM	Research license(mvondavier@ets.org)	GDM family
BUGLIB	Research license(tatsuoka@prodigy.net)	RSM
AHM	Research license(mark.gierl@ualberta.edu)	AHM

5.3 Applications and Simulation studies in CDMs

Whereas earlier CDM literature focused primarily upon theoretical issues such as model development, recently issues in practical applications of the CDMs have been increasing (Huebner, 2010). First, application studies with empirical data have been conducted. Most of the application studies have used tests with mathematical items because mathematical items are simple and easy to analyze Q-matrix (Tatsuoka, 1990; de la Torre and Douglas, 2004; Templin, Henson, & Douglas, 2008; Henson, Templin, & Willse, 2008).

Although empirical studies with math tests demonstrated CDMs are useful in the analysis and interpretation of CDA, the results from the studies cannot be generalizable to various test situations. To overcome this limitation, simulation studies have been conducted. Simulation studies allow researchers to examine various issues in the models and to validate accuracy of diagnostic results from the CDMs across various testing conditions. One of the research areas using simulation method is related to the implementation of CDA for computer adaptive tests. Those studies deal with the issue of how to select items from item banks for diagnostic tests (Huebner, 2010; Cheng, 2009; Xu, Chang, Douglas, 2003). Other topics in the research area of the CDMs include attribute-based scoring in CDMs (Geirl, Cui, & Zhou, 2009), automated test assembly for the CDMs (Finkelman, Kim, & Roussos, 2009), strategies for linking two consecutive diagnostic assessments (Xu & von Davier, 2008), and test construction for cognitive diagnosis (Henson, & Douglas, 2005).

CHAPTER 6

IMPLICATIONS and FUTURE RESEARCHE

CDA and CDMs has some implications in educational practice. Cognitive diagnostic modeling provides a tool that can aid in the development of tailored lesson plans for a student and teachers (Henson, 2009). Through the administration of a series of CDA, students can receive individualized diagnostic reports so that they can make an effort on the area they need to improve. The teachers and the students can have a conference to discuss the kinds of pedagogical actions that they need to take. School administrators and curriculum developers could also receive reports summarizing the students' strengths and weaknesses in tested skills. They could use the information to evaluate the effectiveness of the curriculum innovation. For better administration of CDA in practice, following two issues and future research direction for CDMs are suggested.

6.1 Integrating CDA and Theories of Learning

Huff and Goodman (2007) found that stakeholders want more instructionally relevant assessment results. They indicated that assessments that integrate educational context including cognition, learning, and instruction should be developed in the future. The union of cognitive psychology and psychometrics has been successfully made over

the last several decades. However, little attention was paid to the importance of integrating curriculum, instruction, and assessment to promote student learning using CDA (Huff & Goodman, 2007). For instance, learning theories such as meta-cognition that provide understanding how to use particular strategies for problem solving may help developing effective CDA in educational setting.

6.2. Computer-Based CDA

One more important direction for future CDA is to develop computer-based CDA by using computer technology. E-assessment refers to the use of information technology for any assessment-related activity. Computer-based CDA has advantages over traditional paper-based assessment. To maximize the use of CDA for instructional practice and learning, diagnostic results from any assessments should be sufficiently aligned with the content of the curriculum. Computer-based CDA can yield diagnostic inferences on curriculum specific content student (Gierl & Leighton, 2007). Furthermore, Computer-based CDA can provide immediate score reporting for the teacher and student. Diagnostic feedback in a timely manner with no time lag can be a key factor for the successful assessments. A significant delay between test administration and the reporting of test results is an obstacle to teachers' use of the diagnostic information (Huff & Goodman, 2007). Development of computer-based CDA will overcome this problem of conventional CDA.

Another advantage of computer-based CDA is that it can incorporate alternative item types that are designed to elicit cognitive skills (Jang, 2008). Traditional multiple choice items would have a limitation to provide authentic information about learners' skill competencies. However, various types of tasks can be used in the computer-assisted

CDA. As an example, those tasks are designed for assessing language skills (Jang, 2008): (1) summarizing orally or in writing after listening to a lecture; (2) simulating language use in context; (3) transforming information into a different form (tabulation, graphic representation); or (4) metacognitive reasoning about the appropriateness of language use in a specific context.

The last advantage of the computer-based CDA is that it utilizes various sources of information for diagnosis. The computer-assisted CDA allows the assessment developers and users to consider various sources of information beyond the correctness of the responses to test items. In addition, the computer-assisted CDA can utilize information regarding non-cognitive aspects of learner characteristics in creating diagnostic skill profiles. Individual differences, such as socio-cultural and linguistic background and motivation, may need to be taken into account for the effective diagnosis of learners' skill competencies. Students' self-assessment of skill mastery and problem-solving strategies may enhance their meta-cognitive awareness of the effectiveness of strategy use and facilitate the use of diagnostic feedback for taking remedial actions to change their learning.

Computer-assisted CDA can be used to customize the content of a diagnostic test for each individual student. A sufficiently large item bank with a wide range of skills and task formats will allow teachers to design a diagnostic test that assesses specific skills. The customized tests can be used as tutorials for practice of that students learn.

In conclusion, future direction of CDA implies importance of integrating theories of learning and instruction as well as computer technology into CDA. Such integrations will lead innovative assessments.

6.3. Future Directions of CDMs

Although CDMs are innovative methodologies, they are statistically sophisticated and hard to use in practice. CDMs should be accessible to broad users with simple estimation methods. Developing new tools or refining the algorithms to implement CDMs can be a way to resolve this problem, especially for the models that cannot be analyzed with current software (e.g., RSM).

In addition to the technical issue, comparison among the CDMs under different testing conditions is an important area for future research. For those studies, simulation method should be adapted to set different number of items, examinees, and attributes. The results from the studies can provide insights into the application of the CDMs for potential users. Despite of the importance of such simulation studies, less attention has been paid to them so far. Specifically, there is no simulation studies using RSM and comparing it to other models. Therefore, a comparison of RSM and other models with simulated data is an important area for future research.

CHAPTER 7

SIMULATION STUDY

7.1 Purpose of Study and Simulation Design

The main purpose of this study is to compare diagnoses of skill or attribute possession from three cognitive diagnostic models using simulated data. Even though many CDMs have been developed, comparing diagnosis obtained by Tatsuoka's rule space method (RSM; Tatsuoka, 1983) to those obtained by other CDMs has not been investigated. This problem is due in part to lack of broadly used software for the RSM analysis. In this study, a customized algorism with SAS macros was developed to implement RSM and was applied to simulated data. In addition, the diagnostic results were compared to the results from HO-DINA (de la Torre & Douglas, 2004) and MLTM-D (Embretson & Yang, 2008).

Models to be compared. A primary focus of this study is RSM. Two CDMs, HO-DINA, and MLTM-D have been selected for model comparisons because, like RSM, the diagnoses are based on latent dimensions. That is, the models are non-compensatory cognitive diagnostic models based on IRT models. In RSM, the rule space is constructed by ability and item parameters estimated from an IRT model. In HO-DINA, the probability of mastery is specified by a latent logistic regression model which is similar to IRT models. As a non-compensatory latent trait model, MLTM-D is an extension of IRT models. Finally, all three models provide attribute mastery probabilities and possession of underlying attributes as diagnostic results.

However, there are several differences between the three models in dimensionality of the ability parameters. HO-DINA assumes that mastery of an attribute

is conditionally independent given a single general ability as shown in Equation 4.9. Thus, the proficiency variables in HO-DINA are unidimensional. In contrast, RSM and MLTM-D assumes multidimensionality of the ability parameters. Although classification in traditional RSM is based on one-dimension of θ due to advantage of interpretability and easy computational burden, RSM assumes there are multidimensional latent abilities characterized by observable attributes (Tatsuoka, 2012). Therefore, RSM has a multidimensional characteristic. MLTM-D is also a multidimensional model, but the dimensionality depends on number of component, not attribute level. The other difference between the models is specifications of the cognitive variables in tests. While RSM and HO-DINA requires only attribute structure to specify underlying cognitive process, MLTM-D has two different types of structures: attribute structure matrix (Q-matrix) and component structure matrix (Embretson and Yang, 2012). The details of the three models were described in Chapter 3 and 4.

Data generation design. Since three different CDMs are compared in this study, a single way to generate simulation data only based on one of the CDMs may result in biased true parameters (e.g., attribute patterns or attribute mastery patterns). Consequently, diagnostic estimation from such data is more likely accurate for one model than others. Therefore, in order to control impact of a model on data generation, multiple methods based on different models should be used for data generation. Three different methods of data generation were designed by considering three models – HO-DINA, MLTM-D and RSM. A brief summary of the three methods is presented in Table 4.

Table 4 Summary of Three Methods used in Data Generation

Parameter	HO-DINA	MLTM-D	RSM
Person	Unidimensional a single latent trait (θ_j)	Multidimensional (2 or 3) a vector of latent trait ($\boldsymbol{\theta_{jk}}$)	Multidimensional (6 or 9) a vector of latent trait ($\boldsymbol{\theta_{jk}}$)
Item	location, discrimination (λ_1) slope, difficulty (λ_{0k}) slipping (s_i) guessing (g_i)	intercept for component m (η_{mo}) weight of attribute k on component m (η_{mk})	$p_{jk} \sim U(.50,.99)$ or $p_{jk} \sim U(.01,.49)$

Based on the three models, attribute mastery patterns (a_{jk}), attribute mastery probability (p_{jk}), and item responses (Y_{ij}) of hypothetical examinee *j* were generated. The a_{jk} indicates whether subject *j* has mastered attribute *k* or not. The p_{jk} is a probability that examinee *j* has mastered attribute *k* (e.g., skill or knowledge) given a CDM. These two parameters provide diagnostic information and depend on which model is used. SAS macros were used for all steps of the data generation.

After simulated data sets are generated with three methods based on different models, all three models will be applied to all the data sets. Attribute mastery probabilities ($\hat{p}_{jk}$) and attribute patterns ($\hat{a}_{jk}$) are estimated from the item response patterns for each individual. The accuracy of parameter estimation for each model will be evaluated in terms of average signed bias (ASB), root-mean-square error (RMSE), and correct classification rate (CCR). For relationships of the models, correlation of p_ks across the models and proportion of the same classification will be obtained.

Hypotheses. Comparisons of the diagnostic results will be specifically investigated

in terms of 1) trait dimensionality of diagnostic models as shown in Table 4, 2) the difficulty level of tests, 3) trait level, and 4) number of attributes or components. These variables are manipulated in the simulations. Details will be described in the method section.

First, as described for the dissimilarity of the three models, HO-DINA is a unidimensional model while MLTM-D and RSM are multidimensional models. Therefore, it is hypothesized that recovery of parameter, p_k for cross applications of the models will differ by which model is applied. In other words, for the data set generated by RSM perspective, the attribute mastery probability ($\hat{p}_k$) and the attribute mastery ($\hat{a}_k$) are more accurately diagnosed when MLTM-D is applied than HO-DINA due to the similarity in dimension of RSM and MLTM-D. Likewise, for MLTM-D data, more accurate diagnosis is expected with RSM than HO-DINA. Furthermore, correlations of diagnostic results between RSM and MLTM-D will be higher than between RSM and HO-DINA or MLTM-D and HO-DINA.

Second, the diagnostic results will be compared by level of test difficulty (i.e., easy vs. hard test).The impact of test difficulty level on classification accuracy will be investigated, especially for RSM. In the RSM, classification of a response pattern involves finding close ideal response patterns (i.e., knowledge states). Theoretically, there are only few ideal response patterns with high raw score because all required attributes are necessary for a correct answer in a given knowledge states. When a test is easy, the observed response patterns are likely to have high level of θ, but most of the ideal response patterns have low level of θ. In this situation, the distribution of θ estimated from ideal response patterns will be positively skewed in the rule space

projection. When the distance between ideal points and observed points are not close in most case, it is not possible to do accurate classification. Therefore, level of test difficulty can be an important factor for accurate classification especially in RSM due to the discrepancy between observed response patterns and ideal response patterns. Based on the theoretical characteristic of ideal responses, it is assumed that RSM will have better classification results with difficult test than easy test. It is hypothesized that HO-DINA and MLTM-D will have more accurate diagnostic results than RSM for easy test while RSM will result in similar or better diagnoses for difficult test data. The test difficulty will be adjusted by manipulating attribute difficulty in the data generation. For easy test condition, the item-parameters for each model will be set to have the average p-value of approximately between .6 and .7. For difficult test condition, the item-parameters will be set to have the average p-value of approximately between .4 and .5. P-value is item difficulty index in CTT perspective and means the proportion of examinees that pass an item in a sample.

Third, the diagnostic results will be compared by trait level; $\theta <$ -0.7, -0.7 < θ < 0.7, θ > 0.7 for low, moderate, and high level, respectively. In relation with discrepancy between observed response patterns and ideal response patterns above, when test is easy, RSM is expected to show inaccurate diagnostic results for higher levels of theta than lower or moderate levels of theta, and vice versa. It will be investigate which model is most suitable for which level of ability.

Fourth, the effect of number of attributes on accurate classification will be investigated. The number of attributes was decided by considering the number of knowledge states in latent class models (RSM and HO-DINA). Because 2^k knowledge

states are generated, large number of attributes leads to computational difficulties. On the other hand, it might be hard to do accurate classification with few numbers of attributes. Hence, both manageable number of knowledge states and accurate classification should be considered. In this study, six and nine attributes will be simulated. For RSM applications, the distances between the observed response pattern and ideal response pattern will be more likely to be close in the 9 attribute design than the 6 attribute design because 9 attributes generate a larger number of ideal response patterns than the 6 attributes. Therefore, it is hypothesized that RSM will have more accurate diagnostic results for 9 attribute item design.

This simulation study was designed in item side and person side. A total of 20 replications will be performed for 12 conditions: 2 (item designs) X 2 (level of test difficulty) X 3 (methods of data generation). Since main interest of this study is accuracy of diagnostic results by three models, not parameter estimation for the models, only 20 replications were conducted. The number of examinee for each data set is 1,000. It was elaborated (1) how to construct the Q-matrices (item side) and (2) how to generate simulation data (person side) below. One full replication for 6 and 9 attributes by each of 3 methods has been done, and parts of the data sets will be shown below.

Table 5 Simulation Conditions

Data Generation (3)	Item Design (2)	Test Difficulty (2)
HO-DINA	2 component & 6 attributes	easy
MLTM-D	3 component & 9 attributes	difficult
RSM		

7.2 Method

7.2.1 Item Design

Two different types of item design specified as Q-matrix were used to simulate patterns of attributes in a test. The item designs are intended to apply RSM, DINA and MLTM-D. The item designs are hierarchically structured by including component level as well as attribute level to simulate MLTM-D model. Two Q-matrices are presented in Table 6 and 7, respectively. Components are denoted as 'C', and attributes are denoted as 'A' in the Q-matrices. Each component has multiple attributes. The first item design has two components and six attributes. The other one has three components and nine attributes. The blocks in the tables are constructed by considering possible combinations of the components. For example, items in 'Block 1' in Table 8 involve only component 1, even though they are specified to involve one of the three attributes (A1, A2, or A3). The total number of items is 60 for all item designs. For six attribute design, 66.7 % of the items are single attribute items and 33.3% of the items involve multiple attributes. In contrast, nine attributes design includes 33.3 % of single attribute items and 66.7 % of multiple attribute items.

Table 6 Q-matrix with 6 Attributes

	C1			C2				
Item	A1	A2	A3	A4	A5	A6	Attributes	Block
1	1	0	0	0	0	0	A1	1
2	1	0	0	0	0	0	A1	
3	1	0	0	0	0	0	A1	
4	1	0	0	0	0	0	A1	
5	1	0	0	0	0	0	A1	
6	1	0	0	0	0	0	A1	
7	1	0	0	0	0	0	A1	

Continued

	C1			C2				
Item	A1	A2	A3	A4	A5	A6	Attributes	Block
8	0	1	0	0	0	0	A2	1
9	0	1	0	0	0	0	A2	
10	0	1	0	0	0	0	A2	
11	0	1	0	0	0	0	A2	
12	0	1	0	0	0	0	A2	
13	0	1	0	0	0	0	A2	
14	0	1	0	0	0	0	A2	
15	0	0	1	0	0	0	A3	
16	0	0	1	0	0	0	A3	
17	0	0	1	0	0	0	A3	
18	0	0	1	0	0	0	A3	
19	0	0	1	0	0	0	A3	
20	0	0	1	0	0	0	A3	
21	0	0	0	1	0	0	A4	
22	0	0	0	1	0	0	A4	
23	0	0	0	1	0	0	A4	
24	0	0	0	1	0	0	A4	
25	0	0	0	1	0	0	A4	
26	0	0	0	1	0	0	A4	
27	0	0	0	1	0	0	A4	
28	0	0	0	0	1	0	A5	2
29	0	0	0	0	1	0	A5	
30	0	0	0	0	1	0	A5	
31	0	0	0	0	1	0	A5	
32	0	0	0	0	1	0	A5	
33	0	0	0	0	1	0	A5	
34	0	0	0	0	1	0	A5	
35	0	0	0	0	0	1	A6	
36	0	0	0	0	0	1	A6	
37	0	0	0	0	0	1	A6	
38	0	0	0	0	0	1	A6	
39	0	0	0	0	0	1	A6	
40	0	0	0	0	0	1	A6	
41	1	0	0	1	0	0	A1,A4	
42	1	0	0	0	1	0	A1,A5	
43	1	0	0	0	0	1	A1,A6	
44	1	0	0	1	0	0	A1,A4	

Continued

	C1			C2				
Item	A1	A2	A3	A4	A5	A6	Attributes	Block
45	1	0	0	0	1	0	A1,A5	2
46	1	0	0	0	0	1	A1,A6	
47	1	0	0	1	0	0	A1,A4	
48	0	1	0	0	1	0	A2,A5	3
49	0	1	0	0	0	1	A2,A6	
50	0	1	0	1	0	0	A2,A4	
51	0	1	0	0	1	0	A2,A5	
52	0	1	0	0	0	1	A2,A6	
53	0	1	0	1	0	0	A2,A4	
54	0	1	0	0	1	0	A2,A5	
55	0	0	1	0	0	1	A3,A6	
56	0	0	1	1	0	0	A3,A4	
57	0	0	1	0	1	0	A3,A5	
58	0	0	1	0	0	1	A3,A6	
59	0	0	1	1	0	0	A3,A4	
60	0	0	1	0	1	0	A3,A5	

Table 7 *Q-matrix with 9 Attributes*

	C1			C2			C3				
Item	A1	A2	A3	A4	A5	A6	A7	A8	A9	Attributes	Block
1	1	0	0	0	0	0	0	0	0	A1	1
2	0	1	0	0	0	0	0	0	0	A2	
3	0	0	1	0	0	0	0	0	0	A3	
4	1	0	0	0	0	0	0	0	0	A1	
5	0	1	0	0	0	0	0	0	0	A2	
6	0	0	1	0	0	0	0	0	0	A3	
7	0	0	0	1	0	0	0	0	0	A4	
8	0	0	0	0	1	0	0	0	0	A5	
9	0	0	0	0	0	1	0	0	0	A6	
10	0	0	0	1	0	0	0	0	0	A4	
11	0	0	0	0	1	0	0	0	0	A5	
12	0	0	0	0	0	1	0	0	0	A6	
13	0	0	0	0	0	0	1	0	0	A7	
14	0	0	0	0	0	0	0	1	0	A8	
15	0	0	0	0	0	0	0	0	1	A9	
16	0	0	0	0	0	0	1	0	0	A7	

Continued

	C1			C2			C3				
Item	A1	A2	A3	A4	A5	A6	A7	A8	A9	Attributes	Block
17	0	0	0	0	0	0	0	1	0	A8	1
18	0	0	0	0	0	0	0	0	1	A9	
19	0	0	1	0	0	0	0	0	0	A3	
20	0	0	0	0	0	1	0	0	0	A6	
21	1	0	0	0	0	1	0	0	0	A1,A6	2
22	1	0	0	0	1	0	0	0	0	A1,A5	
23	1	0	0	1	0	0	0	0	0	A1,A4	
24	0	1	0	0	0	1	0	0	0	A2,A6	
25	0	1	0	0	1	0	0	0	0	A2,A5	
26	0	1	0	1	0	0	0	0	0	A2,A4	
27	0	0	1	0	0	1	0	0	0	A3,A6	
28	0	0	1	0	1	0	0	0	0	A3,A5	
29	0	0	1	1	0	0	0	0	0	A3,A4	
30	1	0	0	0	0	1	0	0	0	A1,A6	
31	1	0	0	0	0	0	0	0	1	A1A9	3
32	1	0	0	0	0	0	0	1	0	A1,A8	
33	1	0	0	0	0	0	1	0	0	A1,A7	
34	0	1	0	0	0	0	0	0	1	A2,A9	
35	0	1	0	0	0	0	0	1	0	A2,A8	
36	0	1	0	0	0	0	1	0	0	A2,A7	
37	0	0	1	0	0	0	0	0	1	A3,A9	
38	0	0	1	0	0	0	0	1	0	A3,A8	
39	0	0	1	0	0	0	1	0	0	A3,A7	
40	1	0	0	0	0	0	0	0	1	A1,A9	
41	0	0	0	0	0	1	1	0	0	A6,A7	4
42	0	0	0	0	1	0	1	0	0	A5,A7	
43	0	0	0	1	0	0	1	0	0	A4,A7	
44	0	0	0	0	0	1	0	1	0	A6,A8	
45	0	0	0	0	1	0	0	1	0	A5,A8	
46	0	0	0	1	0	0	0	1	0	A4,A8	
47	0	0	0	0	0	1	0	0	1	A6,A9	
48	0	0	0	0	1	0	0	0	1	A5,A9	
49	0	0	0	1	0	0	0	0	1	A4,A9	
50	0	0	0	0	0	1	1	0	0	A6,A7	
51	0	0	1	1	0	0	0	0	1	A3,A4,A9	5
52	0	1	0	0	1	0	0	1	0	A2,A5,A8	
53	1	0	0	0	0	1	1	0	0	A1,A6,A7	

Continued

	C1			C2			C3				
Item	A1	A2	A3	A4	A5	A6	A7	A8	A9	Attributes	Block
54	0	0	1	0	0	1	0	0	1	A3,A6,A9	5
55	0	1	0	1	0	0	0	1	0	A2,A4,A8	
56	1	0	0	0	1	0	1	0	0	A1,A5,A7	
57	0	0	1	1	0	0	1	0	0	A3,A4,A7	
58	0	1	0	0	1	0	0	0	1	A2,A5,A9	
59	1	0	0	0	0	1	0	1	0	A1,A6,A8	
60	0	0	1	1	0	0	1	0	0	A3,A4,A7	

1 = Measuring, 0 = Not measuring

7.2.2 Data Generation with HO-DINA

To generate simulation data in HO-DINA perspective, the general ability parameter (θ_j), higher-order parameters (λ_{0k}), and item parameters (s_i and g_i) should be determined. These parameters were chosen based on previous simulation studies with HO-DINA (de la Torre & Douglas, 2008; de la Torre & Lee, 2010). First, ability distributions that involve the HO-DINA model should be defined. The θ_js were drawn from a normal distribution with $\mu_\theta = 0$ and $\sigma_\theta = 1$ ($\theta \sim N(0,1)$). The item parameter values for easy test condition are defined in Table 8. Each component consists of three attributes with different levels of difficulty (λ_{0k}), but the discriminating parameters (λ_1) are the same across the attributes within a component.

The higher-order attributes parameters means that the attributes are ordered in terms of difficulty of their mastery (de la Torre & Lee, 2010). In other words, A1 and A4 are easy to master than A3 and A6. Attribute mastery patterns (a_{jk}) were generated by comparing the sampled thetas to the higher-order attribute parameters (λ_{0k}). If $\theta_j > \lambda_{0k}$, a_{jk}=1, otherwise 0. Then, the attribute mastery probabilities were computed by using

Equation (4.9). A sample data was provided in Table 9 with 6 attributes and 15 examinees.

Table 8 HO-DINA Item Parameters for Easy Test Condition

Parameter	2 components X 3 attributes	3 components X 3 attributes
λ_{0k}	{−1.5, -1, -0.5, −1.5, -1, -0.5}	{−1.5, -1, -0.5, −1.5, -1, -0.5, −1.5, -1, -0.5}
λ_1	{1.2, 1.2, 1.2, 0.8, 0.8, 0.8}	{1.2, 1.2, 1.2, 1.0, 1.0, 1.0, 0.8, 0.8, 0.8}

Table 9 Attribute Mastery Probabilities and Attribute Patterns

j	θ_j	p_1	p_2	p_3	p_4	p_5	p_6	a_1	a_2	a_3	a_4	a_5	a_6
1	1.32	0.98	0.18	0.03	0.42	0.18	0.08	1	0	0	0	0	0
2	-0.78	0.59	0.11	0.02	0.25	0.11	0.05	1	0	0	0	0	0
3	-0.30	0.77	0.14	0.03	0.33	0.14	0.06	1	0	0	0	0	0
4	0.25	0.89	0.16	0.03	0.38	0.16	0.07	1	0	0	0	0	0
5	1.15	0.97	0.18	0.03	0.42	0.18	0.08	1	0	0	0	0	0
6	1.22	0.98	0.18	0.03	0.42	0.18	0.08	1	0	0	0	0	0
7	0.06	0.86	0.16	0.03	0.37	0.16	0.07	1	0	0	0	0	0
8	-0.59	0.67	0.12	0.02	0.29	0.12	0.05	1	0	0	0	0	0
9	-0.22	0.79	0.14	0.03	0.34	0.14	0.06	1	0	0	0	0	0
10	1.70	0.99	0.18	0.03	0.42	0.18	0.08	1	0	0	0	0	0
11	-0.87	0.56	0.10	0.02	0.24	0.10	0.04	1	0	0	0	0	0
12	-0.08	0.83	0.15	0.03	0.35	0.15	0.06	1	0	0	0	0	0
13	0.12	0.87	0.16	0.03	0.37	0.16	0.07	1	0	0	0	0	0
14	-0.48	0.71	0.13	0.02	0.30	0.13	0.06	1	0	0	0	0	0
15	0.22	0.89	0.16	0.03	0.38	0.16	0.07	1	0	0	0	0	0

The item responses (Y_{ij}) were generated using attribute mastery patterns (a_{jk}) and the DINA model shown in Equation (4.7). According the DINA model, if all attributes that are required to master an item are mastered by an examinee, the probability passing the item (p_i) is 1-slipping parameter (s_i). If the examinee misses one or more attributes involved in the item, the probability of correcting the item (P_i) is equal to the guessing parameter (g_i). The slip and guessing parameters were fixed to $s = g = 0.2$ for all items. Since many large scale assessment tests are multiple-choice test with five options, 0.2 is a reasonable value as slip and guessing parameter. Random numbers (u) were sampled from uniform distribution between 0 and 1 (U (0,1)). Then, item responses (Y_{ij}) were marked as 1 if P_i is equal or greater than u. If pi is less than u, Y_{ij} were 0. The P_i should be 0.8 or 0.2 because all s_i and g_i parameters are 0.2. Table 10 shows a part of item responses for 8 random items as an example of this step.

Table 10 Item Response Patterns – HO-DINA

j	P_1	P_8	P_{15}	P_{21}	P_{28}	P_{35}	P_{41}	P_{48}	Y_1	Y_8	Y_{15}	Y_{21}	Y_{28}	Y_{35}	Y_{41}	Y_{48}
1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0	0	1	0	1	1	0
2	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	1	1	1	0	1	1	1	1
3	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	1	1	1	0	0	0	0	1
4	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	1	1	1	1	0	1	1	1
5	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0	1	1	1	1	1	1	1
6	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0	1	1	1	1	1	1	1
7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0	1	0	0	1	1	1	1
8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	1	1	1	1	1	0	1	1
9	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	1	1	0	1	1	1	1	0
10	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	1	0	0	0	0	0	0
11	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0	0	1	0	1	1	0

Continued

j	P_1	P_8	P_{15}	P_{21}	P_{28}	P_{35}	P_{41}	P_{48}	Y_1	Y_8	Y_{15}	Y_{21}	Y_{28}	Y_{35}	Y_{41}	Y_{48}
12	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	1	1	1	0	1	1	1	1
13	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	1	1	1	0	0	0	0	1
14	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	1	1	1	1	0	1	1	1
15	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0	1	1	1	1	1	1	1

7.2.3 Data Generation with MLTM-D

For the simulation data generation in terms of MLTM-D, similar method as in DINA was used except separately generating attribute mastery patterns and attribute mastery probabilities by components. The Q-matrix presented in Table 7.2 and Table 7.3 was constructed by reflecting the component structures. Component structure matrices for each item design are shown in Table 11 and Table 12.

Table 11 Component Matrix with 6 Attributes

block	C1 A1, A2, A3	C2 A4, A5, A6	number of items
1	1	0	20
2	0	1	20
3	1	1	20

Table 12 Component Matrix with 9 Attributes

block	C1 A1, A2, A3	C2 A4, A5, A6	C3 A7, A8, A9	number of items
1	1	0	0	10
2	0	1	0	10
3	1	1	0	10
4	1	0	1	10
5	0	1	1	10
6	1	1	1	10

Unlike in HO-DINA, different ability parameters (θ_{jc}) by component (c) were generated from a multivariate normal distribution ($\boldsymbol{\theta} \sim MVN(\mathbf{0}, \sum)$). The number of θ_{jc} is two for two components and 6 attributes design and three for three components and 9 attributes design. It was assumed that each $\theta_c \sim N(0,1)$. The relationship between dimensions will be assumed to be positively correlated each other. In other words, examinees with high level of ability in C1 are more likely to have high level of ability in C2 and C3. Low correlations indicate a high degree of multidimensionality of the test and vice versa (Templin et al., 2008). The same correlations as used in Embretson and Yang (2012)'s simulation study will be adapted to generate the *θc* for three components condition:

$$\sum = \begin{bmatrix} 1 & .3 & .4 \\ .3 & 1 & .5 \\ .4 & .5 & 1 \end{bmatrix}.$$

For two components condition, correlation between C1 and C2 is set to .4:

$$\sum = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}.$$

For two components and six attributes design, the item component difficulty is given as

$$\beta_{im} = \eta_{m0} + \eta_{m1} q_{im1} + \eta_{m2} q_{im2} + \eta_{m3} q_{im3} + \varepsilon_{im}, \qquad (7.1)$$

where η_{mk} is the weight of attribute *k* on component *m*, and η_{mo} is the intercept for component *m*, and q_{imk} indicates binary scores of attribute *k* within component *m* in the item. The q_{imk} can be continuous variables if they are scored for global variables such as complexity level (Embretson and Yang, 2012). Since only Q-matrix was specified in this study, q_{imk} will be 1 or 0. The η_{mo} values were defined to ensure that the overall items

have easy or moderate item difficulty. The η_{mo} and η_{mk} parameter values for each item design are given in Table 13. The variance of ε_{im} was specified as 0.1 for all items. This means that there is a prediction error in testing rather than the attributes perfectly predict item component difficulty. This condition is more close to real testing situation than no error condition.

Table 13 MLTM-D Item Parameters

Parameter	2 components X 3 attributes	3 components X 3 attributes
η_{mo}	η_{10} = -1, η_{20} = -1.5	η_{10} = -2, η_{20} = -1.6, η_{30} = -1.2
η_{mk}	η_{11}=0.3, η_{12} =0.5, η_{13}=0.7	η_{11}=0.3, η_{12} =0.5, η_{13}=0.7
	η_{24}=0.3, η_{25}=0.5, η_{26}=0.7	η_{24}=0.3, η_{25}=0.5, η_{26}=0.7
		η_{37}=0.3, η_{38}=0.5, η_{39}=0.7

MLTM-D model defines component mastery probabilities not attribute mastery probabilities as in RSM and HO-DINA. See Equation 4.27 for details. The P_{ijm} is the probability that examinee *j* applies component *m* correctly on item *i* and it is determined by the ability parameter and the item parameters on component *m*. In the item designs, each component is composed of three attributes, and all items involve only one of the three attributes within components. In other words, each item involves a single attribute nested in each component. Multiple attributes in an item are driven from different components. For example, item 41inclues A1 from component 1 and A4 from component 2. This design makes finding attribute mastery probabilities easy. In Equation 4.27, when number of *k* is 1, P_{ijm} can be simplified as follows by eliminating $\sum_{k=1}^{K}$

$$P_{ijm} = P(Y_{ij} = 1 \mid \theta_{jm}, \underline{q}_{im}, \underline{\eta}_{m}) = \frac{\exp\{-1.7(\theta_{jm} - (\eta_{mk} q_{imk} + \eta_{m0}))\}}{1 + \exp\{-1.7(\theta_{jm} - (\eta_{mk} q_{imk} + \eta_{m0}))\}}. \quad (7.2)$$

In this case, the P_{ijm} can be regarded as the probability mastering attribute *k* within component *m* (i.e., $P_{ijk} = P_{ijm}$). Attribute mastery probabilities for examinee *j* on attribute *k* (P_{jk}) is be obtained from the items involving attribute *k* in the Q-matrix. For example, if item 1 involves only A1, then the attribute mastery probability for A1 will be computed by P_{1j1} (i.e., $P_{j1} = P_{1j1}$). Therefore, attribute mastery probabilities within each component will be computed using Equation (7.2). Next, attribute mastery patters will be obtained by comparing each of the p_{jk} values to the cut-off point of p-value. The cut-off p-value was set to 0.5. This step can be expressed as follows:

$$\begin{cases} a_{jk} = 1, & if\ p_{jk} \geq cut\ point \\ a_{jk} = 0, & otherwise \end{cases}$$

A sample of the data generation with MLTM-D is shown in Table 14.

Table 14 Attribute Mastery Probabilities and Attribute Patterns - MLTM-D

	Component1							Component2						
j	θ_1	p_1	p_2	p_3	a_1	a_2	a_3	θ_2	p_4	p_5	p_6	a_4	a_5	a_6
1	0.10	0.80	0.59	0.87	1	1	1	0.66	0.14	0.22	0.13	0	0	0
2	0.00	0.77	0.54	0.85	1	1	1	2.71	0.85	0.90	0.82	1	1	1
3	0.08	0.79	0.58	0.86	1	1	1	0.47	0.11	0.17	0.09	0	0	0
4	-0.75	0.48	0.25	0.61	0	0	1	0.79	0.17	0.26	0.15	0	0	0
5	1.31	0.97	0.92	0.98	1	1	1	2.63	0.83	0.89	0.80	1	1	1
6	0.43	0.87	0.71	0.92	1	1	1	-0.85	0.01	0.02	0.01	0	0	0
7	1.43	0.97	0.93	0.98	1	1	1	0.74	0.16	0.25	0.14	0	0	0
8	-1.12	0.33	0.15	0.45	0	0	0	-0.55	0.02	0.03	0.02	0	0	0

Continued

	Component1							Component2						
j	θ_1	p_1	p_2	p_3	a_1	a_2	a_3	θ_2	p_4	p_5	p_6	a_4	a_5	a_6
9	0.04	0.78	0.56	0.85	1	1	1	1.81	0.55	0.67	0.50	1	1	1
10	0.26	0.84	0.65	0.90	1	1	1	-0.25	0.04	0.06	0.03	0	0	0
11	-1.48	0.21	0.09	0.31	0	0	0	-0.67	0.02	0.03	0.01	0	0	0
12	1.46	0.98	0.93	0.98	1	1	1	0.87	0.19	0.29	0.17	0	0	0
13	-0.55	0.56	0.32	0.68	1	0	1	1.28	0.33	0.45	0.29	0	0	0
14	0.04	0.78	0.56	0.85	1	1	1	-0.87	0.01	0.02	0.01	0	0	0
15	-0.72	0.49	0.26	0.62	0	0	1	-1.24	0.01	0.01	0.01	0	0	0

To generate item responses for each examinee, the probability that examinee *j* solve item *i*, *P* (Y_{ij} =1), should be computed first. Since MLTD-D is a non-compensatory model, mastering all attributes is required for a correct response. Thus, the probability that examinee *j* solve item *i* (P_{ij}) is the product of probabilities that examinee *j* possess component *m* involved in item *i* (P_{ijm}) as specified in Equation 4.26. For multiple attribute items, the P_{ij} is obtained by multiplying all the $P_{ijm}s$ while the P_{ij} is equal to P_{ijm} for single attribute items. Random number (u) was drawn from uniform distribution ranging from 0 to 1: $u \sim$ U (0,1). Then correct response is determined by comparing each of P_{ijm} to u. This step is simplified as:

$$\begin{cases} Y_{ij} = 1, & if\ P_{ij} = \prod_{k=1}^{K} {p_{jk}}^{q_{ik}} \geq u \sim \mathrm{U}(0,1), \\ Y_{ij} = 0, & otherwise \end{cases}$$

where q_{ik} is a binary variable for the involvement of attribute *k* in item *i* as specified in Q-matrix. Item responses of 8 random items are given in Table 16

Table 16 Item Response Patterns- MLTM-D

j	P_1	P_8	P_{15}	P_{21}	P_{28}	P_{35}	P_{41}	P_{48}	Y_1	Y_8	Y_{15}	Y_{21}	Y_{28}	Y_{35}	Y_{41}	Y_{48}
1	0.80	0.59	0.87	0.14	0.22	0.13	0.12	0.13	1	1	1	0	0	0	0	0
2	0.77	0.54	0.85	0.85	0.90	0.82	0.65	0.49	0	0	0	0	0	0	0	0
3	0.79	0.58	0.86	0.11	0.17	0.09	0.09	0.10	1	1	1	1	1	1	1	1
4	0.48	0.25	0.61	0.17	0.26	0.15	0.08	0.07	0	0	0	0	0	0	0	0
5	0.97	0.92	0.98	0.83	0.89	0.80	0.80	0.82	1	1	1	1	1	1	1	1
6	0.87	0.71	0.92	0.01	0.02	0.01	0.01	0.02	1	1	1	0	0	0	0	0
7	0.97	0.93	0.98	0.16	0.25	0.14	0.16	0.23	1	1	1	0	1	0	0	1
8	0.33	0.15	0.45	0.02	0.03	0.02	0.01	0.01	1	0	1	0	0	0	0	0
9	0.78	0.56	0.85	0.55	0.67	0.50	0.43	0.37	0	0	1	0	0	0	0	0
10	0.84	0.65	0.90	0.04	0.06	0.03	0.03	0.04	0	0	0	0	0	0	0	0
11	0.21	0.09	0.31	0.02	0.03	0.01	0.00	0.00	0	0	0	0	0	0	0	0
12	0.98	0.93	0.98	0.19	0.29	0.17	0.19	0.27	1	1	1	1	1	1	1	1
13	0.56	0.32	0.68	0.33	0.45	0.29	0.18	0.14	0	0	0	0	0	0	0	0
14	0.78	0.56	0.85	0.01	0.02	0.01	0.01	0.01	1	0	1	0	0	0	0	0
15	0.49	0.26	0.62	0.01	0.01	0.01	0.00	0.00	1	0	1	0	0	0	0	0

7.2.4 Data Generation with RSM

As described in Chapter 3, the final goal of classification with RSM is to find attribute mastery probabilities (p_{jk}). To compare accuracy of estimation from RSM to other models, true attribute mastery probabilities should be known. However, it is not possible to directly obtain true p_{jk} from the model because RSM does not specify the probability of mastery as a function of item features and person ability. In RSM, the estimated p_{jk} is based on the distances from the knowledge states in rule space whereas the estimated p_{jk} is obtained from equations with estimated person and item parameters in HO-DINA and MLTM-D. For this reason, the data generation based on RSM was

performed in a different way from the other two methods. The attribute patterns were obtained at the first step, and then the mastery probabilities were generated.

Examinees' mastery or non-mastery (a_{jk}) of an attribute *k* is directly related to examinees' latent trait level on the attribute (θ_{jk}) which is unknown property. When θ_{jk} is known, mastery or non-mastery can be determined by applying cut-off score as proficiency level. The cut-off point is denoted as θ_{cut}. If a θ_{jk} is equal or greater than the θ_{cut}, the examinee *j* will be classified as mastery. The attribute mastery patterns were generated in this way. To generate θ_{jk}s, the θ_{jk} distributions should be considered. It is assumed that the θ_{jk}s are normally distributed with a mean of 0 and a standard deviation of 1 in population. In addition, the θ_{jk}s are assumed to be moderately correlated each other as in MLTM-D. However, the correlations within components (e.g., A1 vs. A2 and A4 vs. A5) are set to be relatively higher than the correlations between different components (e.g., A1 vs. A4) because within component attributes are under the same construct in the item design. The correlation matrices for 6 attributes design are shown in Table 16.

Table 16 Attribute Correlation Matrix for 6 Attribute Q-matrix

	A1	A2	A3	A4	A5	A6
A1	1	.50	.50	.30	.30	.30
A2		1	.50	.30	.30	.30
A3			1	.30	.30	.30
A4				1	.50	.50
A5					1	.50
A6						1

For 9 attributes, the same correlations were used as in 6 attributes design (i.e., .30 or .50). The θ_{jk}s were sampled from a multivariate normal (*MVN*) distribution with the mean vector of $\theta_k(\mathbf{0})$ and the attribute correlation matrix. Since RSM assumes there are multidimensional latent abilities characterized by observable attributes (Tatsuoka, 2012), it is reasonable to sample θ_{jk}s from a multivariate normal distribution.

The sampled θ_{jk}s were used to determine mastery of attribute *k* by comparing them to the cut-off scores. Different cut-off points were applied for different attributes; hence allowing them to vary in difficulty. Low θ_{cut} indicates that high proportion of examinees has mastered the attribute in population, which means the attribute is easy to be mastered by the examinees, and vice versa. For example, if θ_{cut} is -1for A1 and 0 for A2, approximately 84% of the examinees will be classified as mastery for A1while 50% of them will master A2. In this case, A2 is difficult to be mastered by the examinees than A1. For 6 attribute Q-matrix, θ_{cut}= {-1.5, -1.0. -0.5, 1.5, -1.0. -0.5} and for 9 attribute Q-matrix, θ_{cut}= {-1.5, -1.0. -0.5, 1.5, -1.0. -0.5, 1.5, -1.0. -0.5}.

The ability level for each attribute is related to the attribute mastery probabilities. However, in RSM, the attribute mastery probabilities are not a perfect function of θs due to the error of measurement and the probability of misclassifications in diagnosing individual's misconceptions. attribute mastery probabilities were randomly drawn from different uniform distributions for different θ_k levels. The ranges of θ_{jk} and the corresponding uniform distributions are given in Table 17. The correlations between θ_{jk} and p_{jk} were around .90. A sample data was provided in Table 18 with 6 attributes and 10 examinees. For the generation of item responses, same process was used as in MLTM-D because RSM is also a non-compensatory model.

Table 17 Uniform Distributions for Attribute Mastery Probabilities

Attribute	Non-mastery(0)	Mastery(1)
A1, A4,A7	If θ < -1.8, $P \sim U$ (0, 0.249) If -1.8 ≤ θ < -1.2, $P \sim U$ (0.25, 0.499)	If -1.2 ≤ θ < -0.2, $P \sim U$ (0.5, 0.699) If -0.2 ≤ θ < 0.8, $P \sim U$ (0.7, 0.899) If 0.8 ≤ θ, $P \sim U$ (0.9, 1)
A2, A5,A8	If θ < -1.3, $P \sim U$ (0, 0.249) If -1.3 ≤ θ < -0.7, $P \sim U$ (0.25, 0.499)	If -0.7 ≤ θ < 0.3, $P \sim U$ (0.5, 0.699) If 0.3 ≤ θ < 1.3, $P \sim U$ (0.7, 0.899) If 1.3 ≤ θ, $P \sim U$ (0.9, 1)
A3, A6,A9	If θ < -0.8, $P \sim U$ (0, 0.249) If -0.8 ≤ θ < -0.2, $P \sim U$ (0.25, 0.499)	If -0.2 ≤ θ < 0.8, $P \sim U$ (0.5, 0.699) If 0.8 ≤ θ < 1.8, $P \sim U$ (0.7, 0.899) If 1.8 ≤ θ, $P \sim U$ (0.9, 1)

Table 18 Attribute Mastery Probabilities and Attribute Patterns

j	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	a_1	a_2	a_3	a_4	a_5	a_6	p_1	p_2	p_3	p_4	p_5	p_6
1	1.4	0.6	1.3	-0.8	1.2	0.4	1	1	1	0	1	1	0.9	0.8	0.9	0.4	0.7	0.7
2	1.0	-0.2	0.8	1.0	0.8	0.2	1	0	1	1	1	1	0.6	0.4	0.5	1.0	0.6	1.0
3	0.9	1.2	-1.0	1.3	0.4	1.7	1	1	0	1	1	1	0.6	0.8	0.3	0.8	0.8	0.7
4	1.7	0.2	0.8	2.3	1.2	1.1	1	1	1	1	1	1	0.8	0.6	0.7	1.0	1.0	0.9
5	1.0	0.2	-0.3	2.0	2.8	0.5	1	1	0	1	1	1	1.0	0.7	0.2	0.5	0.6	0.7
6	-0.3	0.1	-1.8	-0.8	0.4	-1.0	0	1	0	0	1	0	0.1	0.9	0.2	0.2	0.9	0.5
7	1.1	2.1	-0.1	1.3	0.3	0.6	1	1	0	1	1	1	0.6	0.9	0.1	0.8	0.8	0.7
8	1.3	0.5	-0.3	2.2	1.0	2.4	1	1	0	1	1	1	0.8	0.9	0.3	0.8	0.5	1.0
9	1.3	0.6	-0.2	1.8	1.9	0.8	1	1	0	1	1	1	0.7	0.5	0.2	0.9	0.7	0.7
10	2.7	1.3	2.8	2.0	2.0	0.6	1	1	1	1	1	1	0.7	0.9	0.6	0.8	0.7	0.7

7.2.5 Inspection of Simulated Data

A simple item analysis was conducted with the simulated data sets. P-values and point-biserial correlations from each testing condition were examined. Table 19 and Table 20 shows p-values and point-biserial correlations for a replication of the data sets. P-values are item difficulty index in CTT perspective. Point-biserial correlation coefficient as an item discrimination index represents the correlation between the item score and the total test score. The descriptive statistics shows acceptable range of item characteristics. For easy test condition, item parameters were set to ensure that the mean of p-value for each data set is between .60 and .70. For hard test condition, item responses were generated from items with the mean of p-value is between .40 and .50.

Table 19 Descriptive Statistics of P-value and Point-Biserial Correlation for Easy Test

			P-value			
Method	HO- DINA		MLTM-D		RSM	
Q-matrix	6	9	6	9	6	9
Mean	0.636	0.669	0.634	0.614	0.607	0.611
SD	0.086	0.053	0.126	0.128	0.160	0.128
Max	0.778	0.767	0.862	0.882	0.741	0.771
Min	0.490	0.597	0.439	0.380	0.335	0.301
			Point-biserial Correlation			
Method	HO- DINA		MLTM-D		RSM	
Q-matrix	6	9	6	9	6	9
Mean	0.463	0.513	0.502	0.512	0.483	0.479
SD	0.097	0.118	0.075	0.078	0.051	0.062
Max	0.580	0.604	0.610	0.613	0.673	0.642
Min	0.376	0.345	0.342	0.265	0.343	0.350

Table 20 Descriptive Statistics of P-value and Point-Biserial Correlation for Hard Test

			P-value			
Method	HO- DINA		MLTM-D		RSM	
Q-matrix	6	9	6	9	6	9
Mean	0.427	0.412	0.419	0.417	0.444	0.400
SD	0.075	0.048	0.098	0.116	0.134	0.129
Max	0.668	0.767	0.712	0.680	0.701	0.699
Min	0.210	0.180	0.169	0.170	0.121	0.185

			Point-biserial Correlation			
Method	HO- DINA		MLTM-D		RSM	
Q-matrix	6	9	6	9	6	9
Mean	0.482	0.479	0.565	0.467	0.441	0.401
SD	0.095	0.113	0.087	0.098	0.065	0.061
Max	0.580	0.604	0.610	0.613	0.653	0.642
Min	0.201	0.249	0.312	0.246	0.243	0.220

7.2.6 Cognitive Diagnostic Analysis and Comparisons

The three models were applied all the data set generated with three methods. First of all, in the RSM analysis, the first step is to generating the ideal response patterns. Ideal response patterns are theoretical item score patterns corresponding to each of the attribute patters and used as classification category in RSM.

The ideal response patterns were generated using three different Q-matrices presented above. Constructing an attribute pattern matrix should be proceeded to generate ideal response pattern. This attribute pattern matrix includes all possible combinations among the attributes. For the first Q-matrix, six attribute patterns were included in the matrix. Thus, the total number of attribute patterns is $2^6 = 64$. The third Q-matrix include $512(2^9)$ ideal response patterns. The ideal response patterns were generated from the attribute pattern matrix and the Q-matrix by applying Boolean algebra. Next step of the

RSM analysis is to formulating the rule space was done by calculating the examinees' ability parameters ($\hat{\theta}$) and measures of unusualness (ζ) for both the observed and the ideal response patterns. First, the IRT parameters were estimated with the two parameter logistic model. FlexMIRT (Cai, 2013) was implemented to obtain the Bayesian expected a posteriori (EAP) estimates. The zeta index was calculated by following the equation (3.1). All the processes to obtain the ζ s were performed utilizing SAS macro. In the classification procedure, the Mahalanobis distances were calculated with Equation (3.3). 10 closest distances within each set of ideal points were chosen as a classification category for each response pattern. The attribute mastery probabilities were calculated for each examinee by using Equation (3.4). SAS macro was utilized in all processes of classification.

For HO-DINA, FlexMIRT (Cai, 2013) was used. SAS was also implemented for MLTM-D. Since MLTL-D is a latent trait model, not a latent class model, it provides estimated examinee's ability parameter on a continuous scale for each component or attribute. To compare three models, attribute mastery patterns (α_k) were determined by comparing individual $\hat{\theta}$ to the cutline. If $\hat{\theta} > \theta_{\text{cutline}}$, then $\alpha_k = 1$ and 0, otherwise. Unlike HO-DINA and RSM, MLTM-D only estimates attribute mastery probabilities for each examinee on an item ($\hat{p}_{ijk}$), not for overall test. To obtain overall mastery probability for each attribute on a test ($\hat{p}_{jk}$), mean values of $\hat{p}_{ijk}$ for items that involves the attribute *k* were computed and compared to the results from the other two models.

The diagnostic results from the three models, RSM, DINA, and MLTM-D, were compared. RMSE (root mean square error) and ASB (average signed bias), were used to

check the general accuracy of the parameter estimation ($\hat{p}_k$). These two statistics represent discrepancy between true attribute mastery probabilities (p_{jk}) for each examinee on an attribute and the corresponding estimated probabilities ($\hat{p}_{jk}$) from a CDM. RMSE and ASB can be defined respectively as:

$$\text{RMSE}=\sqrt{\frac{\sum_{j=1}^{N}(p_{jk}-\hat{p}_{jk})^2}{N}} \text{ and } \text{ASB}=\frac{\sum_{j=1}^{N}(p_{jk}-\hat{p}_{jk})}{N},$$

where *N* is the total number of examinees. CCR (correct classification rate) was used to check if a model correctly classifies a response pattern into mastery or non-mastery of an attribute. CCR is the proportion of correct classification in attribute mastery patterns (a_{jk}) for each diagnostic result.

7.3 Results

7.3.1 General Comparison of Three Models

To compare general diagnostic results for each model, comparison statistics were obtained. The results from HO-DINA, MLTM-D, and RSM are shown in Table 21, Table 22, and Table 23, respectively. Means of RMSE, ASB, and CCR at all attributes were presented for each simulation condition in the tables. Each value in the tables is a mean of 6 or 9 attributes. From RMSE and ASB results, it was found that estimated attribute mastery probabilities ($\hat{p}_k$) in RSM are more biased than DINA and MLTM-D for all simulation conditions. Correct classification rates (CCR) also shows better results in HO-DINA and MLTM-D than RSM for most cases.

Table 21 RMSE, ASB, and CCR for HO-DINA

Attributes	Test Difficulty	Data Generation	RMSE	ASB	CCR
6	Easy	HO-DINA	0.230	-0.056	0.972
		MLTM-D	0.250	0.054	0.834
		RSM	0.254	-0.035	0.758
	Hard	HO-DINA	0.251	0.033	0.906
		MLTM-D	0.234	0.010	0.511
		RSM	0.261	-0.026	0.721
9	Easy	HO-DINA	0.212	-0.080	0.991
		MLTM-D	0.250	0.070	0.815
		RSM	0.317	-0.124	0.755
	Hard	HO-DINA	0.244	0.011	0.963
		MLTM-D	0.229	0.008	0.692
		RSM	0.252	0.003	0.686

Table 22 RMSE, ASB, and CCR for MLTM-D

Attributes	Test Difficulty	Data Generation	RMSE	ASB	CCR
6	Easy	HO-DINA	0.172	-0.018	0.840
		MLTM-D	0.262	0.103	0.765
		RSM	0.215	0.020	0.732
	Hard	HO-DINA	0.185	-0.035	0.853
		MLTM-D	0.262	-0.103	0.765
		RSM	0.219	0.025	0.783
9	Easy	HO-DINA	0.218	-0.058	0.802
		MLTM-D	0.277	0.115	0.730
		RSM	0.198	-0.009	0.764
	Hard	HO-DINA	0.183	0.030	0.854
		MLTM-D	0.247	-0.091	0.800
		RSM	0.198	-0.020	0.786

Table 23 RMSE, ASB, and CCR for RSM

Attributes	Test Difficulty	Data Generation	RMSE	ASB	CCR
6	Easy	HO-DINA	0.354	0.151	0.615
		MLTM-D	0.373	0.181	0.561
		RSM	0.343	0.105	0.561
	Hard	HO-DINA	0.363	-0.063	0.603
		MLTM-D	0.329	-0.023	0.622
		RSM	0.352	-0.017	0.512
9	Easy	HO-DINA	0.300	0.165	0.721
		MLTM-D	0.358	0.235	0.627
		RSM	0.327	-0.015	0.649
	Hard	HO-DINA	0.299	-0.045	0.692
		MLTM-D	0.304	0.101	0.662
		RSM	0.305	-0.053	0.604

7.3.2 Number of Attributes

In order to investigate impact of number of attributes on correct classification, marginal means of CCR were obtained for each attribute design and each model. See Table 24 for details. There were no differences in CCR by number of attributes for HO-DINA and MLTM-D. Both 6 and 9 attributes are appropriate for the two models. However, for RSM, CCR was lower in 6 attribute condition than 9 attribute condition.

Table 24 Marginal Means of CCR for Three Models

Model	6 Attributes	9Attributes
HO-DINA	0.756	0.810
MLTM-D	0.779	0.783
RSM	0.579	0.671

7.3.3 Test Difficulty

For RSM applications, it was hypothesized that RSM will show more accurate classification for hard test than easy test. To investigate this hypothesis, marginal means of RMSE, ASB, and CCR in RSM application were presented for each attribute. Table 25 is for 6 attribute design, and Table 26 shows results for 9 attribute design. It was found that level of test difficulty did not have impact on correct classification for both item design. However, all ASBs for easy test are positive while all ASBs for hard test are negative. ASB statistic implies a direction of biases on attribute mastery probabilities. The positive ASB means that $\hat{p}_k$ s were likely to be under-estimated in easy test and 6 attribute condition. The negative ASB means that $\hat{p}_k$ s were likely to be over-estimated in hard test and 6 attribute condition.

Differences on rule space analysis by test difficulty can be observed by projecting the simulated response patterns and the ideal patterns into rule space. See Figure 7 and Figure 8 for 6 attributes condition and 9 attributes condition, respectively. On the rule space, the first dimension represents estimated ability parameters, and the second dimension represent ξ index which is fit index in the RSM. The left plot is rule space projections from easy tests for each data generation method, and the right ones are results from hard tests data. Blue color corresponds to the points of examinees' response patterns, and red color is for ideal response patterns which are knowledge states in the RSM. In 9 attributes condition, ideal response patterns are positively skewed for easy tests data (left side plots).

Table 25 Comparisons by Test Difficulty of RSM Application for 6 Attribute Design

	Easy Test			Hard Test		
	RMSE	ASB	CCR	RMSE	ASB	CCR
A_1	0.288	0.112	0.743	0.361	-0.002	0.621
A_2	0.374	0.172	0.571	0.345	-0.025	0.537
A_3	0.375	0.135	0.540	0.370	-0.050	0.580
A_4	0.327	0.147	0.586	0.331	-0.016	0.573
A_5	0.379	0.167	0.495	0.328	-0.032	0.574
A_6	0.397	0.142	0.540	0.353	-0.081	0.588
M	0.357	0.146	0.579	0.348	-0.034	0.579

Table 26 Comparisons by Test Difficulty of RSM Application for 9 Attribute Design

	Easy Test			Hard Test		
	RMSE	ASB	CCR	RMSE	ASB	CCR
A_1	0.236	0.071	0.861	0.261	0.041	0.757
A_2	0.272	0.091	0.769	0.279	0.013	0.659
A_3	0.366	0.229	0.613	0.313	0.014	0.619
A_4	0.257	0.088	0.797	0.280	0.024	0.645
A_5	0.294	0.097	0.680	0.295	0.001	0.610
A_6	0.444	0.288	0.461	0.323	0.017	0.642
A_7	0.283	0.114	0.753	0.269	0.047	0.713
A_8	0.338	0.118	0.611	0.282	0.028	0.617
A_9	0.467	0.056	0.444	0.420	-0.176	0.610
M	0.328	0.128	0.665	0.302	0.001	0.652

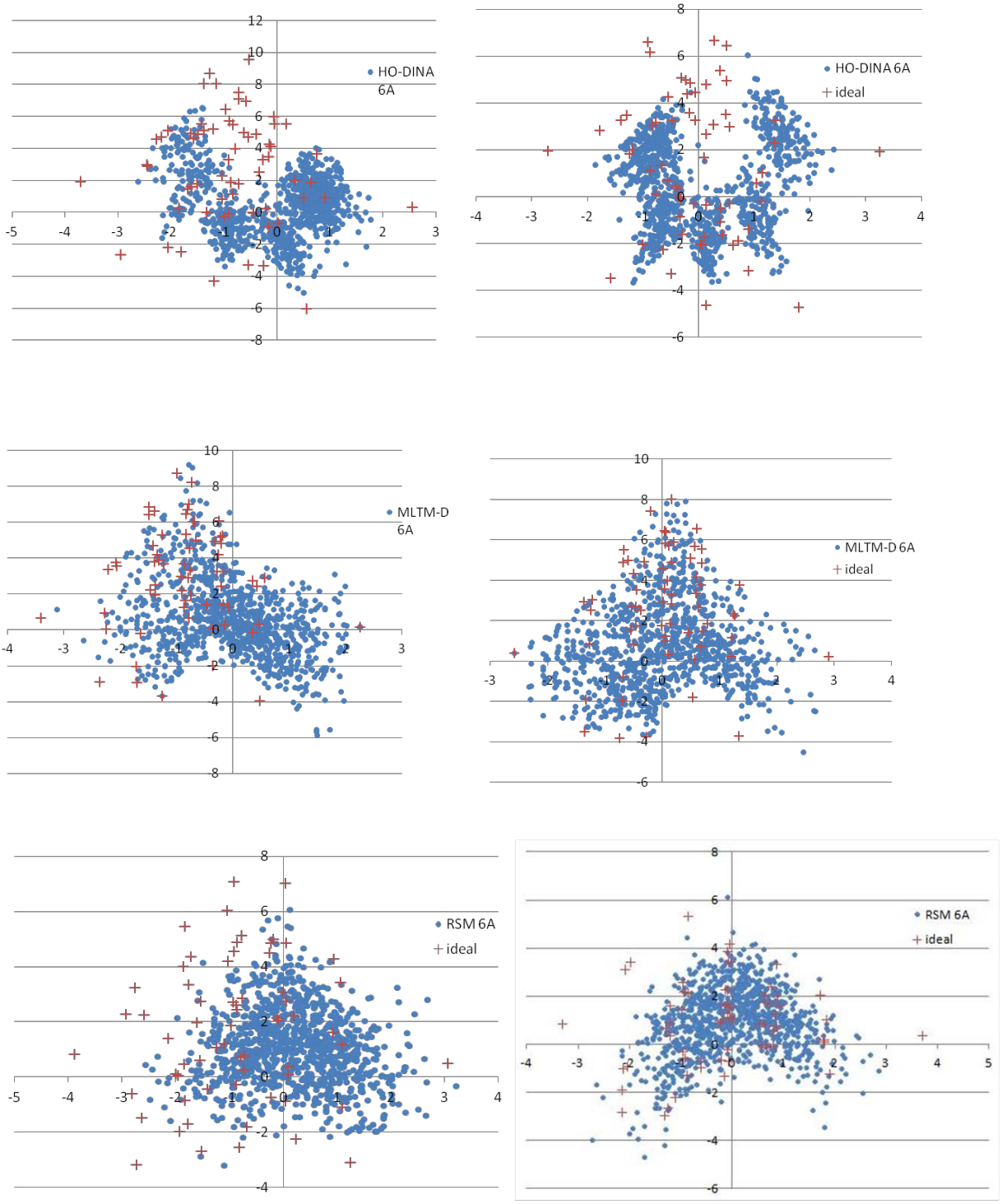

Figure 7. Rule Space for 6 Attribute Condition (left: easy test, right: hard test)

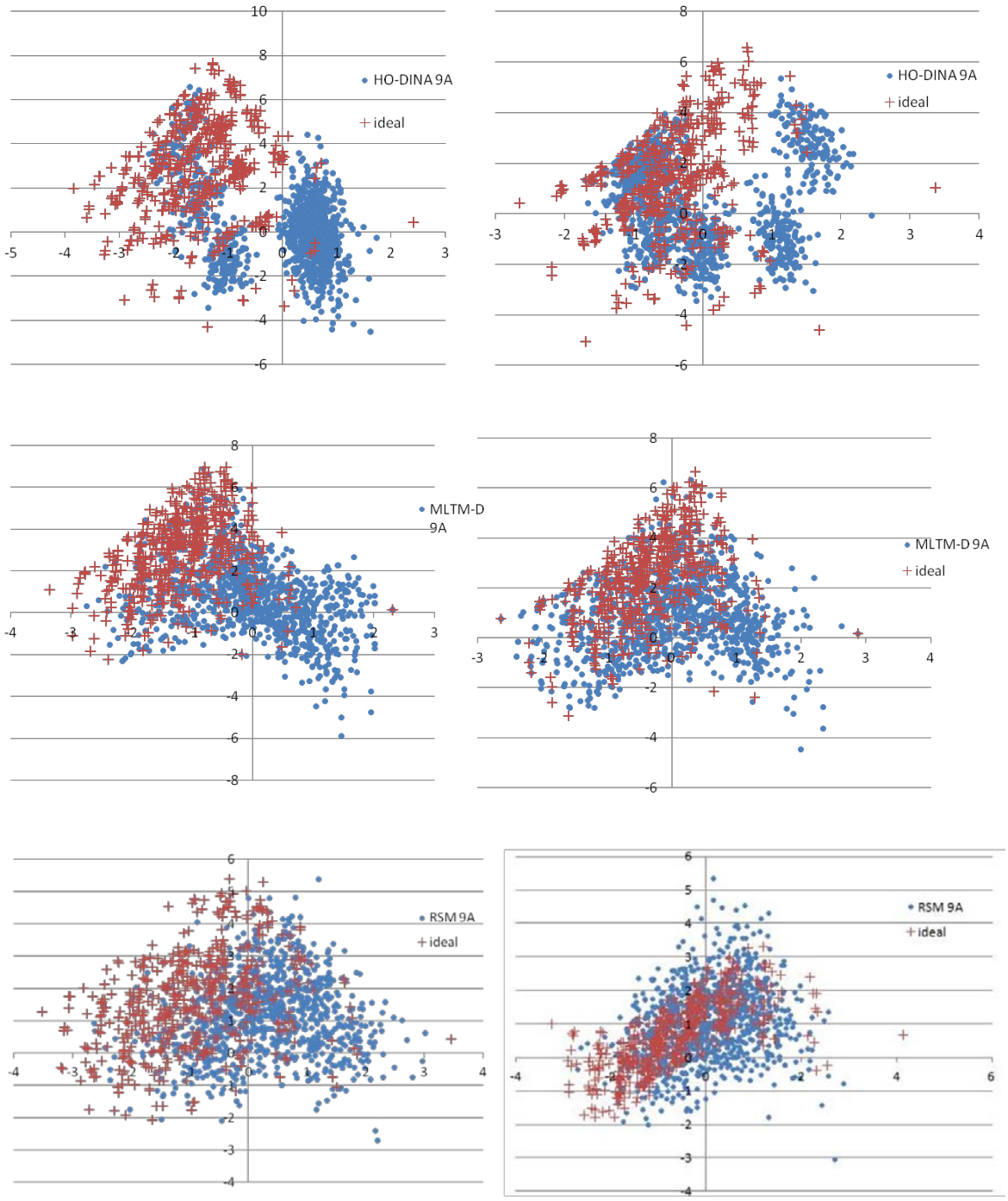

Figure 8. Rule Space for 9 Attribute Condition (left: easy test, right: hard test)

7.3.4 Trait Dimensionality

As described in the first hypothesis, the three models have differences in trait

dimensionality. Based on these differences of the data generation, it is hypothesized that recovery of parameter, p_k for cross applications of the models will differ by which model is applied. However, as shown in Table 21, 22, and 23, the cross application results did not support this hypothesis. The diagnostic results from the three models were not different by the methods of data generation. In other words, dimensionality of ability parameter on the responses data did not influence on accurate classification. HO-DINA and MLTM-D provided better diagnosis than RSM in most conditions. From this finding, average proportions of same classifications of estimated attribute mastery were obtained across the three models for 6 attribute and easy test condition. See Table 27. These values are average of each proportion for 6 attributes. The classification results from MLTM-D and HO-DINA were more similar to each other than RSM. Other tests conditions also provided similar results.

Table 27 Mean Proportion of Same Classification for Attribute Mastery Pattern ($\hat{a}_k$)

Data Generation	HO-DINA vs. MLTM-D	HO-DINA vs. RSM	MLTM-D vs. RMS
HO-DINA	0.837	0.604	0.557
MLTM-D	0.649	0.533	0.472
RSM	0.796	0.533	0.541

7.3.5 Trait Level for RSM

Accuracy of diagnostic results by trait level for RSM was tested. Marginal means of the comparisons statistics across all attributes are given in Table 28. Methods of data

generations were not considered for this investigation. The estimated attribute mastery probabilities from RSM were less biased for low level of trait than for high level of trait.

Table 28 RMSE, ASB, and CCR by Trait Level for RSM

Attributes	Test Difficulty	$\hat{\theta}$	RMSE	ASB	CCR
6	Easy	Low	0.327	-0.171	0.574
		Medium	0.323	0.202	0.633
		High	0.436	0.367	0.618
	Hard	Low	0.403	-0.329	0.625
		Medium	0.302	-0.111	0.609
		High	0.415	0.307	0.566
9	Easy	Low	0.245	-0.056	0.724
		Medium	0.298	0.225	0.670
		High	0.420	0.389	0.630
	Hard	Low	0.239	-0.115	0.769
		Medium	0.296	0.092	0.649
		High	0.389	0.318	0.588

7.3.6 Comparisons of Component Mastery

The item designs used in this study have hierarchical structures (i.e., component and attribute structure). Mastery probabilities at component levels were obtained and compared by the models. See Table 29, Table 30, and Table 30 for the results by three models. MLTM-D has strength when underlying cognitive variables have hierarchical structures. In Table 29 and Table 30, RMSE values were smaller for the data generation based on MLTM-D than the other two methods. In other words, component mastery probabilities from HO-DINA and MLTM-D were less biased when trait dimensions are two or three.

Table 29 Marginal Means of RMSE and ASB for HO-DINA

Components	Test Difficulty	Data Generation	RMSE	ASB
2	Easy	HO-DINA	0.254	0.141
		MLTM-D	0.173	0.168
		RSM	0.263	0.125
	Hard	HO-DINA	0.153	0.053
		MLTM-D	0.129	-0.203
		RSM	0.192	-0.097
3	Easy	HO-DINA	0.280	0.185
		MLTM-D	0.168	0.154
		RSM	0.307	0.045
	Hard	HO-DINA	0.199	-0.085
		MLTM-D	0.174	-0.131
		RSM	0.210	-0.093

Table 30 Marginal Means of RMSE and ASB for MLTM-D

Components	Test Difficulty	Data Generation	RMSE	ASB
2	Easy	HO-DINA	0.254	0.125
		MLTM-D	0.123	-0.134
		RSM	0.143	0.089
	Hard	HO-DINA	0.263	0.064
		MLTM-D	0.129	-0.053
		RSM	0.142	0.047
3	Easy	HO-DINA	0.289	0. 061
		MLTM-D	0.112	0.135
		RSM	0.138	0.019
	Hard	HO-DINA	0.293	-0.048
		MLTM-D	0.124	0.051
		RSM	0.203	-0.083

Table 31 Marginal Means of RMSE and ASB for RSM

Components	Test Difficulty	Data Generation	RMSE	ASB
2	Easy	HO-DINA	0.339	0.179
		MLTM-D	0.373	0.081
		RSM	0.348	-0.105
	Hard	HO-DINA	0.357	0.058
		MLTM-D	0.334	-0.063
		RSM	0.312	-0.089
3	Easy	HO-DINA	0.330	0.147
		MLTM-D	0.339	0.136
		RSM	0.367	-0.059
	Hard	HO-DINA	0.269	0.067
		MLTM-D	0.312	0.199
		RSM	0.309	-0.038

7.4 Discussion

In this study, it was found that HO-DINA and MLTM-D provided more accurate diagnostic results than RSM in all simulation conditions. From this finding, it can be suggested to apply HO-DINA when ability assumed to be unidimensional (i.e., there is only one general ability parameter). When underlying trait is multidimensional, MLTM-D is an appropriate model to be used for diagnosis. Even though MLTM-D is a latent trait model, not latent class model, classification results (i.e., mastery or non-mastery) can be obtained by applying cuff-off level of mastery to the attribute or component mastery probabilities. Determining the cut-off level should be careful and depends on the purpose of diagnostic assessments and level of test difficulty.

In this study, diagnosis from RSM was less accurate than HO-DINA and MLTM-D. However, RSM classifications for 9 attributes condition were better than 6 attributes

condition. This is due in part to the number of ideal response patterns and distances between the actual response patterns and the ideal response patterns. Since, 6 attributes Q-matrix generates a less number of ideal response patterns than 9 attributes Q-matrix (i.e., 64 versus 512 patterns), less number of points are projected into the rule space. This situation might result in inaccurate classification. Thus, with the simulation conditions used in the current study, it cannot be conclude that RSM is not an appropriate model for diagnosis. In future research, other testing conditions for RSM should be investigated with various simulation variables and conditions. For example, the number of attributes involved in items can be a factor that influences on accurate classification. Since RSM is a non-compensatory model, mastery of all required skills is necessary for a correct response. Thus, the number of attributes involved in an item is an important factor of number of correct responses.

Level of test difficulty was not a important factor for accurate diagnosis for HO-DINA and MLTM-D. However, level of test difficulty should be carefully considered when providing attribute mastery probabilities obtained by RSM. In a 6 attributes condition, attribute mastery probabilities were likely to be under-estimated for easy test while they were likely to be over-estimated for the hard tests. Furthermore, the rule space plots in RSM analyses provided less overlap between ideal response points and actual response points. In such case, the diagnostic results from RSM may be less reliable than the situation where ideal and actual response points are homogeneously distributed. Therefore, it is recommended to consider the rule space plot first before obtaining diagnostic results from RSM.

CHAPTER 8

EMPIRICAL STUDY: MATHMETICS TEST DATA

CDMs can be used to demonstrate how well standards-based assessments classify students' level of proficiency. In empirical study, the RSM, HO-DINA, and MLTM-D will be applied to a mathematical test data to estimate real examinees' attribute mastery probability ($\hat{p}_{jk}$) and the mastery pattern (a_k) of each examinee.

8.1 Method

8.1.1 Subjects and Instruments

The state accountability test consisted of 86-items that were administered to all 8th grade students in Kansas. A representative sample of 2993 response patterns was obtained. All items in the test were multiple-choice items with four options and they were dichotomously scored.

Mathematical experts identified that the test for measuring mathematics ability has following four standards: number and computation, algebra, geometry and data. Each standard has two or three benchmarks, and the total number of benchmarks is 10. The benchmarks were used as attributes in the rule space analysis. See Appendix B for details. The Q matrix involving 10 attributes and 86 items was constructed by identifying the required skills for getting an item correct on the test. The Q-matrix is available in Appendix A.

8.1.2 Cognitive Diagnostic Analysis

Three diagnostic models, RSM, HO-DINA, and MLTM-D, were applied to the empirical data by following the same process as in the previous simulation study.

8.2 Results

Descriptive statistics of diagnostic results will be provided. Attribute mastery probabilities of 10 attributes are given in Table 32. Comparisons for classifications are available in Table 33. As found in the previous simulation study, HO-DINA and MLTM-D provided similar classification.

Table 32 Descriptive Statistics of Attribute Mastery Probabilities (N=2993)

	HO-DINA			MLTM-D			RSM		
	Min	*Max*	*M*	*Min*	*Max*	*M*	*Min*	*Max*	*M*
Raw Score	22	86	60.68 (15.12)	22	86	60.68 (15.12)	22	86	60.68 (15.12)
$\hat{p}_1$	0.000	0.999	0.533 (0.412)	0.003	0.999	0.585 (0.397)	0.000	1.000	0.702 (0.397)
$\hat{p}_2$	0.001	1.000	0.645 (0.402)	0.005	0.941	0.590 (0.278)	0.000	1.000	0.789 (0.278)
$\hat{p}_3$	0.000	1.000	0.651 (0.410)	0.070	0.958	0.619 (0.294)	0.000	1.000	0.706 (0.294)
$\hat{p}_4$	0.000	1.000	0.436 (0.435)	0.160	0.929	0.526 (0.408)	0.000	1.000	0.601 (0.408)
$\hat{p}_5$	0.000	1.000	0.535 (0.416)	0.030	0.967	0.645 (0.402)	0.000	1.000	0.575 (0.402)
$\hat{p}_6$	0.006	1.000	0.637 (0.407)	0.078	0.923	0.596 (0.308)	0.000	1.000	0.606 (0.308)
$\hat{p}_7$	0.001	1.000	0.638 (0.404)	0.038	0.926	0.613 (0.270)	0.000	1.000	0.687 (0.270)
$\hat{p}_8$	0.166	1.000	0.804 (0.287)	0.129	0.979	0.756 (0.254)	0.000	1.000	0.416 (0.254)
$\hat{p}_9$	0.000	1.000	0.432 (0.427)	0.180	0.920	0.530 (0.436)	0.000	1.000	0.440 (0.436)
$\hat{p}_{10}$	0.001	1.000	0.643 (0.400)	0.020	0.959	0.683 (0.345)	0.000	1.000	0.383 (0.345)

Table 33 Proportion of Same Classification for Examinee Attribute Mastery Pattern ($\hat{a}_k$) (N=2993)

	HO-DINA vs. MLTM-D	HO-DINA vs. RSM	MLTM-D vs. RSM
a_1	0.89	0.49	0.55
a_2	0.78	0.52	0.49
a_3	0.91	0.41	0.54
a_4	0.94	0.38	0.47
a_5	0.71	0.58	0.60
a_6	0.68	0.35	0.46
a_7	0.64	0.56	0.39
a_8	0.71	0.42	0.58
a_9	0.69	0.47	0.35
a_{10}	0.80	0.58	0.46

Fit Statistics for HO-DINA, MLTM-D, and 2PL IRT model are presented in Table 34. Since RSM does not provide fit statistics, goodness of fit was tested with 2PL IRT model. The results indicated that HO-DINA did not fit as well as the MLTM-D.

Table 34 Goodness of Fit

	HO-DINA	MLTM-D	2PL IRT
-2lnL	245,935	242,549	238,783
AIC	246,301	243,818	239,127

AIC: Akaike Information Criterion

Descriptive statistics for estimated theta and zeta index in RSM are given in Table 35. $\hat{\theta}_{actual}$ and ζ_{actual} were obtained from students' actual response patterns, and were obtained from ideal response patterns. The plot in the rule space is presented in Figure 9. An example of diagnostic results by RSM is available in Table 36.

Table 35 Descriptive Statistics for Estimated Theta and Zeta in RSM (*N*=2993 for actual; *N*=1024 for ideal)

	Ideal Score	Actual Score	$\hat{\theta}_{actual}$	ζ_{actual}	$\hat{\theta}_{ideal}$	ζ_{ideal}
Min	0.00	22.00	-1.94	-2.81	-3.98	0.28
Max	86.00	86.00	2.62	9.00	1.11	16.10
Mean	37.03	60.68	0.00	1.02	-1.98	7.11
SD	14.64	15.12	0.97	1.60	0.73	2.75

Table 36 RSM Classification Results for a Student ($\hat{\theta}_{=}$ -1.76)

Knowledge State	Theta	Zeta	Mahalanobis distance	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
0362	-1.3338	3.8370	0.0024	0	0	1	0	0	1	1	1	1	0
0653	-1.3334	3.8326	0.0026	1	1	1	0	0	1	1	0	1	1
0177	-1.3343	3.8798	0.0030	1	1	0	1	0	1	1	1	0	0
0261	-1.4513	3.8302	0.0031	1	1	0	1	1	0	1	0	1	0
0657	-1.3334	3.8843	0.0033	1	1	0	1	0	1	1	0	1	1
0659	-0.1984	3.9088	0.0034	1	1	1	1	0	1	1	0	1	1
0608	-1.7080	3.8457	0.0036	0	0	1	1	0	1	0	0	1	1
0763	-1.1366	3.7905	0.0054	1	0	0	1	1	1	1	1	1	1
0376	-1.3338	3.9098	0.0057	1	0	1	0	1	1	1	1	1	0
0307	-2.2212	3.8528	0.0060	1	0	0	1	1	0	0	1	1	0

Attribute Mastery Probabilities = **[0.8 0.5 0.5 0.7 0.4 0.8 0.8 0.5 0.9 0.5]**

Table 36 shows an example of classification results for an examinee. This student is classified into 10 knowledge states by finding 10 closest ideal points on the rule space.

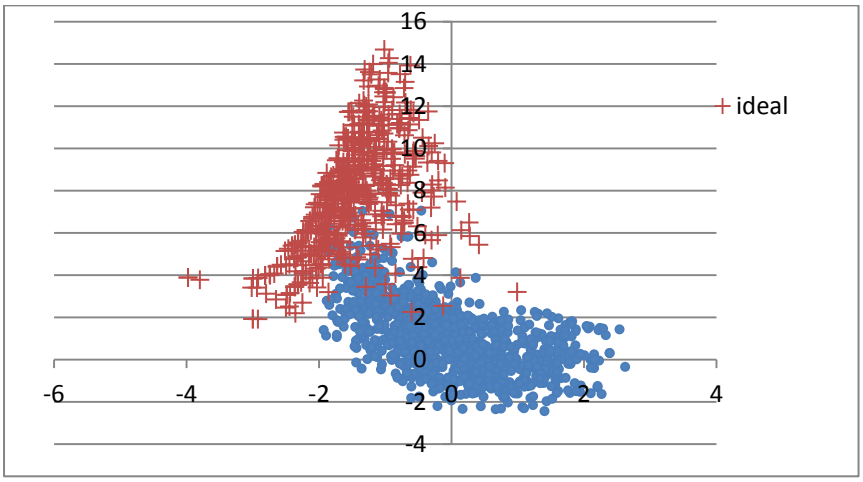

Figure 9. Rule Space Plot (red: ideal, blue: student's pattern)

8.3 Discussion

In the empirical study, the three models were applied to a standard-based assessment test data. As found in the previous simulation study, the diagnostic results (attribute mastery probabilities and the attribute mastery patterns) from HO-DINA and MLTM-D were more similar each other than those from RSM. RSM is a less appropriate model to be applied to this data than MLTM-D and HO-DINA. This test has easy level of test difficulty because the mean of raw scores was 60.68 out of 86 items and the mean p-value of the items was moderately high (M =0.70). However, the attribute structure of this test is not simple. The test involves 10 attributes and 28% of the items requires multiple attributes for a correct answer. The ideal response patterns based on the Q-matrix involves only few response patterns (knowledge states) with high level of theta,

which result in in accurate classification for students with high level of theta. Since high proportion of the examinees have high level of theta, overall diagnostic results from RSM could be less reliable than the other models.

When considering the characteristics of the Q-matrix, MLTM-D seems to be the most appropriate model for the data because the Q-matrix involevs 4 standards and each standards has 2 or 3 benchmark. Therefore, multidimensionality of the underlying traits should be assumed for diagnosis. Furthermore, order relationships in attribute difficulty are not assumed for the Q-matrix. However, if diagnoses are conducted by standards, HO-DINA is a good model to be applied. Even though sub indicators for each benchmart (attribute) were not presented in the cognitive skill structure, there are several sub indicators in the test blue print.

In future studies, other CDMS can be applied to this data. Since only compensatory models were applied, comparing to non-compensatory models (i.e., GDM or LCDM) will be necessary to investigate if this type of test is appropriate for a compensatory model or non-compensatory model.

APPENDIX A:Q-MARIX FOR KANSAS STATE 8th GRADE MATHEMATICS TEST

Item	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
1	1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	1
3	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	1	0
5	0	1	1	0	0	0	0	0	0	0
6	0	1	1	0	0	0	0	0	0	0
7	0	1	1	0	0	0	0	0	0	0
8	0	1	1	0	0	0	0	0	0	0
9	0	0	0	0	0	1	0	1	0	0
10	0	0	0	0	0	1	0	1	0	0
11	0	0	0	0	0	1	0	1	0	0
12	0	0	0	0	0	0	1	0	0	0
13	0	0	0	0	0	0	1	0	0	0
14	0	0	0	0	0	0	1	0	0	0
15	0	0	0	0	1	0	0	0	0	0
16	0	0	0	0	1	0	0	0	0	0
17	0	1	0	0	0	0	0	0	0	0
18	0	1	0	0	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	0	0
20	0	0	0	1	0	0	0	0	0	0
21	0	0	0	1	0	0	0	0	0	0
22	0	0	0	1	0	0	0	0	0	0
23	0	0	0	1	0	0	0	0	0	0
24	0	0	1	0	0	0	0	0	0	0
25	0	0	1	0	0	0	0	0	0	0
26	0	0	1	0	0	0	0	0	0	0
27	0	0	0	0	0	1	0	1	0	0
28	0	0	1	0	0	0	0	0	0	0
29	0	0	1	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	1
31	0	0	0	0	0	0	0	0	0	1
32	0	0	0	0	0	0	0	0	0	1
33	0	0	0	0	0	0	0	0	0	1
34	0	0	0	0	0	0	0	0	0	1
35	0	0	1	0	0	1	0	0	0	0

Continued

Item	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
36	0	0	1	0	0	1	0	0	0	0
37	0	0	1	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	1	0	0
39	0	0	0	0	0	1	0	1	0	0
40	0	0	0	0	0	1	0	1	0	0
41	0	0	0	0	0	1	0	1	0	0
42	0	1	0	0	0	0	0	0	0	0
43	1	0	0	0	0	0	0	0	0	0
44	0	1	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	1	0	0	0
46	0	1	0	0	0	0	0	0	0	0
47	0	0	1	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	1	0
49	0	0	0	0	0	0	0	0	1	0
50	0	0	0	0	0	0	0	0	1	0
51	0	0	0	0	0	0	0	0	1	0
52	0	0	0	0	1	0	0	0	0	0
53	0	0	0	1	0	1	0	0	0	0
54	0	0	0	1	0	1	0	0	0	0
55	1	0	0	0	0	0	0	0	0	0
56	0	0	0	0	0	0	1	0	0	0
57	0	0	0	0	1	0	0	0	0	0
58	0	0	0	0	1	0	0	0	0	0
59	0	0	0	0	1	0	0	0	0	0
60	0	0	0	0	1	0	0	0	0	0
61	1	0	0	0	0	0	0	0	0	0
62	0	0	0	0	0	1	1	0	0	0
63	0	0	0	0	0	1	1	0	0	0
64	0	0	0	0	0	1	1	0	0	0
65	0	0	0	0	0	1	1	0	0	0
66	0	0	0	0	0	1	1	0	0	0
67	0	0	1	0	0	0	0	0	0	0
68	0	0	1	0	0	0	0	0	0	0
69	0	0	0	1	0	0	0	0	0	0
70	0	0	1	0	0	0	0	0	0	0
71	0	0	0	1	0	1	0	0	0	0
72	0	0	0	1	0	0	0	0	0	0
73	0	1	0	0	0	0	0	0	0	0

Continued

Item	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
74	0	0	0	0	0	0	0	0	1	0
75	0	0	0	0	0	0	0	0	1	0
76	0	0	0	0	0	0	0	0	1	0
77	0	0	0	1	0	0	0	0	0	0
78	0	0	0	1	0	0	0	0	0	0
79	0	0	0	0	0	0	0	0	1	0
80	0	0	0	0	0	0	0	0	1	0
81	0	0	0	0	0	0	0	0	1	0
82	0	0	1	1	0	0	0	0	0	0
83	0	0	0	1	0	0	0	0	0	0
84	0	0	0	1	0	1	0	0	0	0
85	0	0	1	1	0	0	0	0	0	0
86	0	0	0	1	0	0	0	0	0	0

APPENDIX B: ATTRIBUTES FOR THE KANSAS 8th GRADE MATHENATICS TEST

Attribute	Attribute Description
	I. Number and Computation
A1	1. Number Sense: The student demonstrates number sense for real numbers and simple algebraic expressions in a variety of situations.
A2	2. Number System and Their Properties: The student demonstrates an understanding of the real number system; recognizes, applies, and explains their properties; and extends these properties to algebraic expressions.
A3	3. Computation: The student models, performs, and explains computation with rational numbers, the irrational number pi, and algebraic expressions in a variety of situations.
	II. Algebra
A4	4. Variable, Equations, and Inequalities: The student uses variables, symbols, real numbers, and algebraic expressions to solve equations and inequalities in a variety of situations.
A5	5. Functions: The student recognizes, describes, and analyzes constant, linear, and nonlinear relationships in a variety of situations.
A6	6. Models: The student generates and uses mathematical models to represent and justify mathematical relationships found in a variety of situations.
	III. Geometry
A7	7. Geometric Figures and Their Properties: The student recognizes geometric figures and compares their properties in a variety of situations.
A8	8. Geometry form an Algebraic Perspective: The student uses an algebraic perspective to examine the geometry of two dimensional figures in a variety of situations.
	IV. Data
A9	9. Probability: The student applies the concepts of probability to draw conclusions, generate convincing arguments, and make predictions and decisions including the use of concrete objects in a variety of situations.
A10	10. Statistics: The student collects, organizes, displays, explains, and interprets numerical (rational) and non-numerical data sets in a variety of situations.

REFERENCES

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223-237.

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

Bechtoldt, H. P. (1959). Construct validity: A Critique. *American Psychologist, 14*(10), 619-629.

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

Cai, L. (2013). flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2006, April). *A person-fi statistic for the attribute hierarchy method: The hierarchy consistency index.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

Davier, M., DiBello, L., & Yamamoto, K. (2008). Reporting test outcomes using models for cognitive diagnosis. In J. Hartig, E. Klieme, D. Leutner, J. Hartig, E. Klieme, D. Leutner (Eds.) *Assessment of competencies in educational contexts* (pp. 151-174). Ashland, OH US: Hogrefe & Huber Publishers.

DeCarlo, 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(1), 8-26.

De La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353. de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement,* 46(4), 450-469.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1993). *Unified cognitive/psychometric Diagnosis foundation.* Manuscript submitted for publication.

DiBello, L., Roussos, L. A., & Stout, W. F. (2007). Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, 26*, 979-1030.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471-516). Washington, DC: American Council on Education.

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

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

Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics.*New York: Academic Press.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56(3)*, 495-516.

Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). New York: Plenum Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods* 3, 300-396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika* 64, 407-433.

Embretson, S. E., & Yang, X. (2006). Multicomponent Latent Trait Models for Complex Tasks. *Journal Of Applied Measurement, 7*(3), 335-350.

Embretson, S. E. & Yang, X. (2013). A multicomponent latent trait model for diagnosis. Psychometrika, 78, 14-36.

Finkelman, M., Kim, W., & Roussos, L. (2009). Automated test assembly for cognitive diagnostic models using a genetic algorithm. *Journal of Educational Measurement, 46(3)*, 273-292.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica* 37, 359-374.

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

Issues and practice, 19, 23-44.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement, 44*(4), 325-340.

Gierl, M.J., Leighton, J.P. & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills. In Leighton, J.P. & Gierl, M.J. *Cognitive diagnostic assessment for education: Theories and applications* (pp. 242-274). Cambridge, MA: Cambridge University Press.

Gierl, M. J., Alves, C., & Majeau, R. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, *10*(4), 318-341.

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

Glas C. W. (2003, March) On the Structure of Educational Assessments: A Comment From a Psychometric Perspective. *Measurement: Interdisciplinary Research And Perspectives* [serial online], *1*(1):76-79.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika,* 61, 215-231.

Gorin, J. S. (2009). Diagnostic classification models: Are they necessary? Commentary on Rupp and Templin (2008). *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 30-33.

Griffin, P. & Nix, P. (1991). Educational Assessment and reporting, A new approach. New South Wales: Harcourt Brace Jovanovich.

Hartz, S. (2002) *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301-323.

Henson, R. A. (2009). Diagnostic classification models: Thoughts and future directions. *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 34-36.

Henson, R., & Douglas, J. (2005). Test Construction for Cognitive Diagnosis. *Applied Psychological Measurement*, *29*(4), 262-277.

Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, 32(4), 275-288.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2), 191-210.

Huebner, A., (2010). An Overview of Recent Developments in Cognitive Diagnostic Computer Adaptive Assessments. *Practical Assessment, Research & Evaluation*, 15(3). Available online:http://pareonline.net/getvn.asp?v=15&n=3.

Hunt, E. (1995). Where and When to Represent Students This Way and That Way; An Evaluation of Approaches to Diagnostic Assessment. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 411-429). Hillsdale, NJ: Erlbaum

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.

Jiao, H. (2009). Diagnostic classification models: Which one should I use?. *Measurement: Interdisciplinary Research and Perspectives*, 7(1), 65-67.

Lai, C. P. and P. Griffin (2001). Linking Cognitive Psychology and Item Response Models. Annual Conference of the Australian Association for Research in Education, Perth.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Leighton, J. P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*, 6-15.

Leighton, J. P. (2008). Where's the psychology? A commentary on "unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art.'. *Measurement: Interdisciplinary Research and Perspectives*, 6(4), 272-275.

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

Leighton, J. P., & Gierl, M. J. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees'

thinking processes. Educational Measurement: Issues and Practice, 26, 3-16.

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

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4*, 333– 369.

Luecht, R. M. (2006a, May). Engineering the test: From principled item design to automated test assembly. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Luecht, R. M. (2006b, September). Assessment engineering: An emerging discipline. Paper presented in the Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada.

Luecht, R. M. (2007, February). Assessment engineering workshop. Presented at Association of Test Publishers Conference. Palm Spring, CA.

Luecht, R. M. (2008, February). *Assessment engineering*. Session paper at the Assessment Engineering: Moving from Theory to Practice, Coordinated panel presentation at the Annual Meeting of the Association of Test Publishers, Dallas, TX.

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). Scalability and the development of useful diagnostic scales. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Marshall, S. P. (1995). Some Suggestions for Alternative Assessments. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 431-453). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

Messick, S. **(**1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J. (1993). Foundations of a new test theory. In Frederikson, N., Mislevy, R. J.,&Bejar, I. (Eds.), *Test theory for a new generation of tests*. Hillsdale, New Jersey:Lawrence: Erlbaum Associates, Publishers.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K .B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 437-446). San Francisco: Morgan Kaufmann.

Mislevy, R.J., Almond, R.G., & Lukas, J. (2004). A brief introduction to evidence-centered design. (CSE Technical Report 632). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST).

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

Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

No Child Left Behind Act of 2001, Pub.L. No. 107-110, 115 Stat. 1435 (2002)

Roussos, L. A., DiBello, L. V., Stout, W. F., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. Leighton., & M. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education.* New York, NY: Cambridge University Press.

Roussos, L. A., DiBello, L. V., Henson, R. A., Jang, E. E., & Templin, J. L. (2008). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. Embretson & J. Roberts (Eds.), *New directions in psychological measurement with model-based approaches.* Washington, DC: American Psychological Association.

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

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

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response models. In J. P. Leighton, M. J. Gierl, J. P. Leighton, M. J. Gierl (Eds.) , Cognitive diagnostic assessment for education: Theory and applications (pp. 205-241). New York, NY US: Cambridge University Press.

Sympson J., B. (1978). A model for testing with multidimensional items. In Weiss DJ (ed) Proceedings of the 1977 *Computerized Adaptive Testing Conference*, University Of Minnesota, Minneapolis

Snow, R.E.(1989). Towards assessment of cognitive and cognitive structures in learning. *Educational Researcher, 18(9),* 8-14.

Snow, R. E., & Lohman, D. F. (1989). Implication 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.

Snow, R. E., & Lohman, D. F. (1993). Cognitive psychology, new test design, and new test theory: An introduction. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 1-18). Hillsdale, NJ: Erlbaum.

Templin, J. (2006). *CDM user's guide.* Kansas, NE: University of Kansas.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305.

Toulmin, S. E. (1958), *The Uses of Argument*, Cambridge University Press, Cambridge.

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. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*(1) 55-73.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diadnosis. In N. Fredrickson, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds). *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum

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

Tatsuoka, K. K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method.* New York: Routledge Academic

Tatsuoka, M. M. (1971). *Multivariate analysis*. New York: John Wiley.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*(3), 195-211.

van der Linden, W. J. (2005). Linear models for optimal test design. New York: Springer-Verlag.

van der Maas, H. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, 118(2), 339-356.

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

von Davier, M. (2005b). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: ETS.

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Simharay (Eds.), Handbook of statistics: Psychometrics (Vol. 26, pp. 643–661). Amsterdam: Elsevier.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287-307.

von Davier, M. (2008b). *The Mixture General Diagnostic Model.* In G. R. Hancock & K. M. Samuelson (Eds.), Advances in Latent Variable Mixture Models. Information Age Publishing.

von Davier, M. (2009). Some Notes on the Reinvention of Latent Structure Models as Diagnostic Classification Models. *Measurement: Interdisciplinary Research & Perspectives,7*(1),67-74.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*(1) 8-28.

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Simharay (Eds.), Handbook of statistics: Psychometrics (Vol. 26, pp. 643–661). Amsterdam: Elsevier.

von Davier, M., & Yamamoto, K. (2007). Mixture distribution Rasch models and Hybrid Rasch models. In M. von Davier & C. H. Carstensen (Eds.),

Multivariate and mixture distribution Rasch models (pp. 99–115). New York: Springer.

Wilson, M. (2005). *Constructing measures: An item response theory modeling*. Mahwah, NJ: Erlbaum.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck, & M. Wilson (Eds.), *Explanatory Item Response Models: A generalized linear and nonlinear approach.* New York: Springer.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479-494.

Xu, X., Chang, H., Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.

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

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models* (ETS Research Report RR-89-41). Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service No. ED 310161).

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). Design and analysis in a cognitive assessment (ETS RR-03-32). Princeton: NJ: ETS.

Zhou, Jiawen (2009, October). A Review of Assessment Engineering Principles with Select Applications to the Certified Public Accountant Examination, AICPA Technical Reports

ACHKNOWLEDGEMENTS

First of all, I wish to thank my advisor, Dr. Susan Embretson, for her patience and for her supports throughout my doctoral study. Her guidance has allowed me to continue my dissertation work. I would like to express my sincere gratitude to my committee members, Dr. Tofighi, Dr. Thomas, Dr. Parsons and Dr. Templin. I thank them for providing me valuable feedback and encouragement.

My parents deserve the special recognition. I am indebted to my parents' love and support. I hope to continue to make them proud with my accomplishments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

SUMMARY

Cognitive diagnostic assessment (CDA) is a new theoretical framework that is designed to integrate cognitive psychology into measurement theories. The main purpose of CDA is to provide examinees with diagnostic information while traditional psychometric approaches focus on how latent variables are accurately measured. Many cognitive diagnostic models (CDM) have been developed for CDA. Three cognitive diagnostic models- namely the rule space method (RSM), the high-order deterministic inputs, noisy 'and' gate (HO-DINA) model, and the multidimensional latent trait model for diagnosis (MLTM-D) model were compared using simulated data and empirical data. For the simulation study, three methods of data generation are proposed. Each method was designed based on one of the three models. A total of 12 conditions was involved in the simulation study: 2 item designs X 2 level of test X 3 methods of data generation. The diagnostic results were compared by level of test difficulty, level of ability estimates, and level of dimensionality. The effect of number of attributes on accurate classification was also investigated. For the empirical study, a mathematics test data was used and the diagnostic results were compared.