Cognitive Diagnostic Model, a Simulated-Based Study: Understanding Compensatory Reparameterized Unified Model (CRUM)

Roofia Galeshi

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
In
Educational Research and Evaluation

Gary E. Skaggs, Chair
Penny L. Burge
Kusum Singh
Min Sun

(September 24, 2012)
Blacksburg, VA

Keywords: cognitive diagnostic model, CRUM, relative fit indices, model selection

Cognitive Diagnostic Model, a Simulated-Based Study: Understanding Compensatory Reparameterized Unified Model (CRUM)

Roofia Galeshi

Abstract

A recent trend in education has been toward formative assessments to enable teachers, parents, and administrators assist students succeed. Cognitive diagnostic modeling (CDM) has the potential to provide valuable information for stakeholders to assist students identify their skill deficiency in specific academic subjects. Cognitive diagnosis models are mainly viewed as a family of latent class confirmatory probabilistic models. These models allow the mapping of students' skill profiles/academic ability. Using a complex simulation studies, the methodological issues in one of the existing cognitive models, referred to as compensatory reparameterized unified model (CRUM) under the log-linear model family of CDM, was investigated. In order for practitioners to implement these models, their item parameter recovery and examinees' classifications need to be studied in detail. A series of complex simulated data were generated for investigation with the following designs: three attributes with seven items, three attributes with thirty five items, four attributes with fifteen items, and five attributes with thirty one items. Each dataset was generated with observations of: 50, 100, 500, 1,000, 5,000, and 10,000 examinees. The first manuscript is the report of the investigation of how accurately CRUM could recover item parameters and classify examinees under true QMattrix specification and various research designs. The results suggested that the test length with regards to number of attributes and sample size affects the item parameter recovery and examinees classification accuracy. The second manuscript is the report of the investigation of the sensitivity of relative fit indices in detecting misfit for over- and opposite-Q-Matrix misspecifications. The relative fit indices under investigation were Akaike information criterion (AIC), Bayesian information criterion (BIC), and sample size adjusted Bayesian information criterion (ssaBIC). The results suggested that the CRUM can be a robust model given the consideration to the observation number and item/attribute combinations. The findings of this dissertation fill some of the existing gaps in the methodological issues regarding cognitive models' applicability and generalizability. It helps practitioners design tests in CDM framework in order to attain reliable and valid results.

Dedication

This dissertation is dedicated to Mehdi, Ali, and Cameron Setareh who have been my constant source of inspiration. They have given me the drive, enthusiasm, and determination to persist on the task.

Acknowledgement

I have heard this phrase before but did not realize its depth and truth until now, that writing a dissertation is a never ending process; there is always room for improvement. However, I would have never been able to finish this work without the help of my advisor and committee chairman, Dr. Gary Skaggs, who has helped me throughout this difficult task and has introduced me to this new exciting field of psychometrics.

Also my committee members, Dr. Penny Burge, whom has helped me during the writing process, Drs. Elizabeth Creamer, Kusum Singh, and Min Sun for their constructive comments and suggestions which helped me to express my thoughts more clearly.

I would also like to express my appreciation to my father, Hassanali Galeshi, for his emotional and financial support for as long as I remember. He has taught me to be strong, persistent, and never give up.

My mother passed away on 2004 but I would like to express my love, my gratitude and my admiration to my mother; she was the light of my life since I was born. I would like to thank my sister, Rozita Galeshi, and my nephew, Hamidreza Teimoory, for their support and encouragements along the way.

This study was partially supported by the National Science Foundation under the grant No. 1109429. This support is gratefully acknowledged. Any opinions, findings and conclusions expressed in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

Table of Content

List of Tables

List of Abbreviation

AIC	Akaike information criterion
BIC	Bayesian information criterion
CAT	Computerized adaptive testing
CDM	Cognitive diagnostic models
CFA	Confirmatory factor analysis
CFI	Comparative fit index
DCM	Cognitive diagnostic classification models
DINA	Deterministic inputs noisy and-gate
DINO	Deterministic inputs noisy or-gate
GDM	General diagnostic model
IC	Information indices
IRT	Item response theory
LC	Latent class
LLTM	Latent logistic test model
MAD	Mean absolute deviation
MCMC	Markov chain Monte Carlo
NAEP	National assessment of educational progress
NIDA	Noisy-input deterministic-and-gate
NIDO	Noisy-input deterministic-or-gate
PISA	Programme for international student assessment
PSAT	Scholastic Assessment Test
RMSE	Root mean squared error

RMSEA	Root mean square error of approximation
RUM	Reparameterized unified model
ssaBIC	Sample size adjusted Bayesian information criterion
TIMSS	Trend in international mathematics and science study
TOEFL	Test of English as a foreign language

Chapter 1

Diagnostic Models

Cognitive diagnostic models (CDMs) are latent structure models similar to the flexible family of generalized linear mixed models introduced by von Davier (2005). CDM statistically classifies examinees by assigning each respondent into pre-assigned latent classes.

According to Leighton and Gierl (2007), “cognitive diagnostic assessment (CDA) is designed to measure specific knowledge and processing skills in students so as to provide information about their cognitive strength and weaknesses” (p. 3). On the other hand, the well-established item response theory (IRT) mainly estimates a person's location on a single latent trait scale and does not provide specific diagnosis on individual skill mastery. IRT models (de Ayala, 2009; Embretson & Reise, 2000) classically report a single score for a ranked comparison of examinees. The results provide limited information about each student and preclude tailoring programs to individual students' needs. Traditional models have also been criticized for not providing detailed information about each student’s strengths and weaknesses in a specific academic subject (Snow & Lohman, 1989; Leighton & Gierl, 2007). It is increasingly evident that these new methodologies, CDMs, will become a critical component of testing.

Little research has been done in the understanding and application of CDMs, more specifically in the compensatory reparameterized unified model (CRUM). Although, some researchers have paid attention to this new methodology (de la Torre & Douglas, 2004; Henson, Templin, & Willse, 2009; Kunina-Habenicht, 2010; Rupp, Templin, & Henson, 2010), little is known about these models’ behavior and application under various sample sizes, attributes, and test length, specifically in CRUM approach.

While, CDMs have gained the focus and attention of researchers, a review of the model literature reveals relatively little theoretical work in the area of parameter recovery, examinee classification, and relative fit indices. Because of the unique potential of these new methodologies, investigating their behavior in various research designs can fill the existing gap and help to establish a general guideline for practitioners.

From several methodological issues existing in the CDM analysis one serious area of concern is the lack of in depth existing research on fit statistics. Currently there are no established fit statistics available for CDMs. Hence, most researchers' approach toward model fit is the comparison of nested models with relative model fit indices such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) (Rupp, Templin, & Henson, 2010).

Chi-square statistics (χ^2) are commonly used as fit indices in statistical and measurement analysis. The chi-square statistic compares the estimated versus observed data and compares the P-value with the chi-squared distribution. A global fit statistic of chi-square is difficult to apply to CDMs models based on their large number of latent class variables and hence sparse data. Agresti and Finlay (1997) stated that the chi-square distribution can be the sampling distribution only if the sample is roughly larger than five observations per cell.

This restriction is very difficult to impossible to achieve for most cognitive models due to the exponential progression of latent classes. Complexity of CDM models increases exponentially with the addition of each attribute to the model. For K number of attributes 2^{K} possible latent class patterns exist. For instance, a diagnostic model with three attributes and 100 observations has 2^3=8 latent classes, an average of 100/8= 12.5 observation per latent class. On the other hand, a diagnostic distribution with 10 attributes and 100 observation will create 2^{10}= 1024 latent classes and an average of 100/1024=0.1 observation per cell, a lot less than 5 per cell observation requirement for χ^2 estimation, hence, it will be impossible to accurately estimate model fit with χ^2 (Rupp, Templin, & Henson, 2010).

To avoid the encounter of sparse cells, researchers have suggested the use of relative fit indices (de la Torre & Douglas, 2008; Rupp, Templi, & Henson, 2010). The assumption would be that the data fits a CDM model and the task is to find the model that fits the data best. This approach allows for the Q-Matrix validation. de la Torre and Douglas (2008) used this technique to investigate the problem of Q-Matrix misspecification for his DINA Model. One important aspect of cognitive models is the Q-Matrix. A fundamental issue in developing these models is the accurate specification of such a matrix. The Q-Matrix can be viewed as a loading matrix similar to those used in factor analysis with elements that identifies the existence or lack of specific skill in each item.

Examining empirical data is of great advantage for understanding the performance of the technique in its natural environment, but so far there have only been a few educational tests with CDM structure in mind, not available to public and limited in sample sizes and skills (Kunina-Hbenicht, 2010). This scarcity of available data leads most researchers to utilize existing test and retrofit the models which causes convergence problem and lack of clear evidence of fit. To appreciate the potential benefits of these models in educational assessments and to eliminate the difficulties in retrofitting, this dissertation will focus on the data collected from simulation of CDM model to be able to truly examine the performance of the model.

Rationale and Justification

Cognitive Diagnostic Models are fairly new methodologies with few researchers investigating their performance under various conditions. There are a small number of research studies available on these models, yet none cover all available CDMs and all possible practical research designs.

A major point of difference between this study and others is the chosen model as well as its wide range of sample sizes, item length, and number of attributes. Haberman, von Davier, and Lee’s (2008) work on cognitive models was limited to four skills, Kunina-Habenicht’s (2010) study was confounded to two sample sizes of 1,000 versus 10,000 with three and five attributes, Rupp and Templin’s (2008) study of DINA model included only four attributes, and Templin and Henson’s (2006) investigation of DINA model consisted of 40 items, seven attributes with 3,000 respondents.

Research on relative fit indices, to the knowledge of this author, has been limited to three existing manuscripts. The first study was conducted by de la Torre and Douglas (2008) on comparing the DINA and NIDA models using a simulated test of 20 items and five attributes. The relative fit indices suggested that DINA was superior to NIDA. The second study was conducted by Choi, Templin, and Cohen (2010) with 40 items, four attributes, and four different sample sizes, suggesting that the relative fit indices were able to detect misfit for CRUM for datasets of 200 and more examinees. The third study was conducted by Kunina-Habenicht (2010) on performance of relative fit indices for three and five attributes with 1,000 and 10,000 examinees; her study revealed that the relative fit indices was able to detect misfit for both sample sizes.

This researcher aims to fill the existing gap in the literature of cognitive models. I investigate the theoretical potential and difficulties of one of the existing cognitive models, CRUM, for various sample sizes of 50, 100, 500, 1,000, 5,000, and 10,000; various attribute/item combinations of three attributes with seven and 35 items, four attributes with 15 items, and five attributes with 31 items using simulated datasets. This document includes two manuscripts; each study makes an important contribution to the psychometric literature.

The first manuscript is the report of examination of the CRUM parameter recovery under five various sample sizes and four various attribute/item combinations. It also investigated the accuracy of examinee's classification into their accurate latent classes. The second manuscript describes the examination of the sensitivity of the relative fit indices (AIC, BIC, and sample-size-adjusted BIC) in CRUM. It aimed to offer theoretical and practical guidelines for estimation of CRUMs, mostly for educational applications.

Research Questions

The main goal of this dissertation was to investigate CRUM's ability in recovering item parameters under various research designs. It aimed at exploring the relative fit indices sensitivity in detecting misfit for CRUM under various designs. We have investigated several important underexplored issues regarding CRUM by the means of simulated data. the following research questions are addressed in this study:

1. How accurately does the model recover its parameters under various test length/ Q-Matrix and sample size conditions when dichotomous data is generated from a CRUM?
2. How accurately does the model classify students into their true class of attribute mastery under various test length/ Q-Matrix and sample size conditions when dichotomous data is generated from a CRUM?
3. How accurately do the three relative fit indices (AIC, BIC, sample size adjusted BIC) detect the fit of the model under various *test length/Q-Matrix* and *sample size conditions* when the dichotomous data is generated from a CRUM?

4. How do different relative fit indices such as AIC, BIC, and sample-adjusted BIC compare in detecting misfit under various conditions of model misspecification when the dichotomous data is generated from a CRUM?

Organization of the Dissertation

This dissertation consisted of five chapters with the following organization. Chapter 1 is a background of the studies, a justification for the studies, a synopsis of the two manuscripts, and summary of the results was demonstrated. Chapter 2 is a comprehensive literature review on existing studies on cognitive diagnostic models. Chapter 3 is the first manuscript on CRUM parameter recovery. Chapter 4 is the second manuscript on relative fit indices sensitivity in detecting misfit for datasets with CRUM specifications. Chapter 5 is a discussion chapter; it discusses the implications of these studies, their limitations, and recognizes areas for future research.

Chapter 2

Literature Review

During the last decade, cognitive diagnostic modeling has gained a great deal of interest as a promising area of psychometric research. Its aim is to identify students' mastery status of a pre-defined group of skills, or attributes, in order to provide them with a detailed description of their strengths and weaknesses in the measured subject matter. The purpose of this literature review is to provide an understanding of cognitive diagnostic models (CDMs) and to describe their existing statistical approaches. The main focus is on one of these models, the compensatory reparametrized unified model (CRUM) that is based on a log-linear approach. In the literature as well as this dissertation, the presence of an attribute is referred to as *mastery* of that attribute and the absence of an attribute is referred to as *nonmastery*.

This chapter is organized in the following manner: a basic overview and introduction to CDMs, the importance of cognitive models, an overview of methodologies similar to CDM, its main structure, main categories of CDMs, a brief discussion of the six CDM core, a more detailed review of CRUM with an hypothetical example, and CDM's application and issues, research gap in CRUM.

Introduction to Cognitive Models

While teachers have various tools for judging students' achievement, when it comes to large scale assessments CDM is an option for helping teachers and parents to focus on specific strengthening skills. Huff and Goodman's (2007) survey of K-12 teachers revealed that a great majority of teachers (93%) believed that it is very important to collect diagnostic information using an assessment tool. It is difficult for teachers to understand and infer information regarding their students' performance from the statistical scales available from item response theory (IRT) models. Cognitive models, on the other hand, can serve as an instructional tool by providing detailed attribute assessment of student's performance and placing them into their appropriate class mastery (Linn, 1990).

Rupp, Templin and Henson (2010) defined the cognitive diagnostic models as "[s]tatistical models with discrete latent variables that are used to classify respondents into one of several distinct latent classes associated with individual attribute profiles" (p. 319). CDMs are probabilistic/ psychometric models combining cognitive science with psychometrics analysis that provide rich information for the purpose of training (de la Torre & Douglas, 2008; Roussos, Dibello, Stout, Hartz, Henson, & Templin, 2007). CDM is intended to identify the cognitive processes behind a student's response to a particular test item. This analytical family has been new to the field of psychometrics and offers great potential in *evaluating* test scores and student's thought process (Lighton & Gierl, 2007).

The Importance of Cognitive Models in Test Analysis and Evaluation

Interest in the growth of cognitive modeling has been largely due to the United States educational leader's call for more formative assessments made by the No Child Left Behind Act of 2001 (No Child Left Behind, 2002). Mislevy (1995) explained that current test theory has been effective in selecting students most likely to succeed in a particular educational institution or program but has yet to be effective in helping students succeed. This assertion contrasts with the main goal of standardized educational assessment, which is to collect information on student learning so as to improve it. A long-standing view on the testing field has focused on a summative approach, measuring achievement, or selecting individuals, rather than testing for diagnostic purposes (Linn, 1989; Norris, Macnab, & Phillips, 2007). CDM can connect a formative approach to measuring student's achievement with standardized testing.

One of the most important applications of cognitive models, Embretson (1991) argued, is its ability to offer informative feedback that is easy to understand to parents, teachers and students, which can then be used to improve instruction (Embretson, 1998). She further argues that the process of examining test scores needs to be based on a substantive theory of the learner's process toward the task, a tool in validating the results of cognitive diagnostic assessments. The traditional IRT and classical test theory (CTT) analysis provides information about a students' achievement, not their profile of attributes explaining their level of achievement (Rupp & Mislevy, 2007). Similarly de la Torre and Douglas (2008) stated that the main CDM goal is to create a diagnostic assessment tool based on a person's performance on a test.

Moreover, Harre (1970) specified that the depth of traditional scientific investigations is limited to classification, to using the direct observable behaviors, and to assessing how such events might concur. He and others further argued that an event can be explained by finding its causal mechanism (Huff and Goodman, 2007). Cognitive models have the potential to map these causal mechanisms, that is, students' cognitive strengths and weaknesses, whereas the more traditional test scoring methods—increasingly recognized as informative but not adequate—do not (Mislevy, Almond, & Lukas, 2004).

Cognitive models detail characteristics of student attribute mastery by analyzing students' responses to test items, measuring mastery of latent trait/attribute rather than assessing general ability. Any examinee who has mastered those required attributes is more likely to answer test items correctly than one who either did not master any of the required attributes or one who has mastered only a few attributes. The measure of mastery is a question of probability. For example, an item on a math test might require both multiplication and division for an accurate response;

however, a student who has mastered both attributes has a higher probability of responding accurately than a student who has mastered only one or none of the attributes. Generally, examinee *r* has a (*1 x K*) vector, attribute mastery pattern, that locates said examinee into his or her appropriate latent class.

Methodologies Similar to Cognitive Models

Cognitive diagnostic models (CDMs) are a family of confirmatory probabilistic models with categorical latent variables (Kunina-Habenicht, 2010). Other techniques such as cluster analysis and factor analysis can provide similar information about students' classification. Although cluster analysis is a useful method for data reduction, one promoting interpretability of data based on observed measures, it is not a hypothesis driven or a statistical model (McCutcheon, 1987). Grouping the participants into "clusters" does not provide any statistical information such as probability of membership to a group.

Factor analysis, on the other hand, is a statistical technique that can be used for analysis of latent variables. However, factor analysis is largely used for latent continuous variables which have continuous distributions. Attributes are dichotomous variables and do not have a normal distribution (Bruin, 2006; Muthen & Muthen, 2010). So, unlike factor analysis, CDM analyzes discrete—and not continuous—variables.

It is easier to separate factor analysis and cluster analysis from CDM since the difference is clear, but even within methods intended to provide diagnostic assessment, there are methods, such as rule space methodology (RS) and Attribute Hierarchy Method (AHM), that are quite different statistically. They are latent variable psychometric models and both approach diagnostic assessment differently from CDM's insofar as they are not log-linear approaches to data analysis. However, for comparison purposes, they will be discussed briefly below.

Rule space methodology introduced by Tatsuoka (1983) is a pattern analysis of student's item responses on tests. Its purpose is to classify students according to their profile of strengths and weaknesses relative to their underlying attributes/constructs. Application of RS analysis includes several phases: (1) defining attributes, (2) assigning attributes to items, (3) finding classification space identified by student's IRT ability (Ɵ), and (4) *Zeta (ξ)* the measure of "unusualness of response". RS classifies students based on their item responses while accounting for these unusual responses that students might demonstrate despite their repertoire of attributes.

The attribute hierarchy method (AHM) is a variation on RS by considering a hierarchy of attributes (Leighton, Gierl, & Hunka, 2004). AHM assumes that attributes do not live in isolation; rather mastery of one could be dependent on the mastery of another. This concept stands in relief to most existing cognitive models including RS that assumes attributes are functioning independently from each other. AHM, on the other hand, emphasizes that a student's performance on a test depends on a set of hierarchically ordered sets of attributes proficiencies.

The cognitive models under the focus of this literature review sets apart from those existing latent variable psychometric models in their approach and outcome. CDMs are fully statistical modeling approaches with parameters and predictors unlike RS and AHM that are not statistical models with parameters to estimate.

CDMs in log-linear format are inspired from Fischer's (1973) logistic latent test model (LLTM). He argued that experiential evidence illustrates that the psychological complexity of a math problem cannot be summarized in an ordinary statistical description of the item difficulty. He noticed there is no significant increase of complexity in an item when the same operation occurs repeatedly within that item. He proposed the weighted matrix to indicate whether the attribute is required by the item for correct response. To measure the probability of the correct response in conjunction with item difficulty he employed the logistic model:

$$logit(P_{vi}) = \sum_{j} f_{ij}\eta_j + \varepsilon_v + c \qquad (1)$$

$$logit\ (P_{vi}) = \ln(\frac{P_i}{(1-P_i)}) \qquad (2)$$

In the equation above, *i* represents the item, *j* represents the attribute, and *v* represents the examinee. Moreover, P_{vi} represents the probability that subject *v* correctly responses to item *i*; f_{ij} represents the weight factor, in other words, the binary matrix of operations (attributes) in item *I*; η_j is the item difficulty; ε_v represents the ability of subject *v;* and *c* represents the constant.

In summary, CDMs, as being used here, are confirmatory probabilistic item response (IRT) models with categorical latent variables that allows for multiple structures to describe the observed responses. Personal attributes similar to personality characteristics are considered to be latent constructs. These attributes, latent variables/constructs, can be measured through observable item responses.

Main Structure of CDM's

So far a general overview of CDM followed by the importance of CDM methodology and how it differs from like approaches were described, chiefly, cluster analysis, factor analysis, RS, and AHM. Cognitive diagnostic models share several basic structures such as attributes, Q-Matrix, and latent class profile which will be discussed below.

Attributes

Attributes have been referred to with various names: The literature in the CDM field labels attributes as skills, latent traits, or latent characteristics (Rupp, Templin, & Henson, 2010).

While the term skills or attributes can be used interchangeability to refer to examinees' measured latent characteristic, for consistency, this literature review uses attributes.

Similar to ability in item response theory, skills or attributes are latent traits of educational characteristics and are not usually directly observable. These latent characteristics can only be measured indirectly. In other words, the latent variables explain the relationship between the observed variables and student responses to the test items (McCutcheon, 1987).

Each test item has its own pre-assigned vector of 1's and 0's indicating which skills are measured by that specific item (here 0's indicate the absence of skills). For instance, on a test that measures five skills, an item requiring skills one and four for an accurate response will have a [1, 0, 0, 1, 0] item vector in the model. These vectors will create the rows of a Q-Matrix. Observed responses can be related to the skills by the means of a loading matrix referred to as the Q-Matrix.

Q-Matrix

The theory behind the Q-Matrix was first introduced by Tatsuoka (1983) and was inspired by the work of Fischer's (1973) logistic latent test model (LLTM). The presence or absences of attributes are the elements of the Q-Matrix. Each element of the Q-Matrix is designated by q_{ik}, indicating whether mastery of attribute k is required for correct response to item i. If k corresponds with a correct response to item i, then $q_{ik} = 1$, otherwise $q_{ik} = 0$. The Q-Matrix is a n x k matrix with items in the row and attributes in the column. The Q-Matrix identifies which attributes needed to be mastered to answer each item correctly.

$$q_i = \begin{cases} 1\, if\ the\ attribute\ is\ needed\ for\ item\ i \\ 0\ \ if\ the\ attribute\ is\ not\ needed\ for\ item\ i \end{cases}$$

An example of a Q-Matrix looks as follows for a three-item test measuring two attributes:

$$QMatrix = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The above Q-Matrix indicates item one requires both attributes one and two, item two requires only the first attribute, and item three requires only the second attribute for an accurate response.

The most fundamental part of CDM estimation is identifying the necessary skills to accurately respond to each item on the test (Tatsuoka, 1983). Loading matrix specification is one of the central structures of CDM models (Rupp, Templin, Henson, 2010). The building of a Q-Matrix involves theory and application in praxis. The theory is based on cognitive processes that are hypothesized to operate in an academic or non-academic task. These cognitive processes support the application of assigning numeric values to the Q-Matrix (Rupp, Templin, & Henson, 2010). Hence, it is essential to construct the Q-Matrix accurately. The misspecification of the Q-Matrix could have profound effect on the accuracy of the estimation of the CDM models.

Latent Class Attribute Profile Mastery

The last common structure of CDMs is the latent class attributes mastery profile. The main purpose of these models is to create a discrete attribute mastery profile for each examinee based on their responses to test items. These attribute profile consists of a series of probability of mastery classifications (Rupp, Templin, & Henson, 2010). Attribute profile mastery indicates the examinees skill in a dichotomous array created by the posterior probability of latent class membership to which each examinee belongs. If *K* represents the number of skills being measured by a test, where each specific skill is shown by *k* and can take on a value of 0 or 1, a respondent *r* can have a vector indicating his/her skill mastery profile as follows:

$$C_r = (\alpha_{r1}, \alpha_{r2}, \dots \alpha_{rk}) \tag{3}$$

A latent class with an array of 1's indicates the examinee has mastered all the skills involved in the test, while an array of 0's indicate non- mastery of the skills required by the test items. The attributes have exponential relationship to the latent classes, meaning as the number of attributes *K* increases linearly, the number of possible latent classes increases to 2^{K}. The latent class membership assumes a causal relationship between the examinees' responses and the unobserved latent skill mastery. For example, on a test with three attributes, subtraction, addition, and multiplication, an examinee could be estimated to have one of $2^3 = 8$ latent classes.

The Main Categories of CDM Families

CDMs are classified into two categories: non-compensatory and compensatory (Jiang, DiBello, & Stout, 1996). Rupp, Templin, and Henson (2010) have a comprehensive review of these models for those interested in more detailed description, however, for this literature review, the main distinction in CDM families is on the manner in which the latent attributes are joined together for a correct response to a test item. CDMs are subcategorized into six core models, three as non-compensatory and three as compensatory models (Rupp, Templin, & Henson, 2010). The suitability of the model relates to the measured latent characteristics. What follows is a brief description of each category.

Non-compensatory model. A non-compensatory or conjunctive model indicates that mastery of one attribute *cannot compensate for non-mastery of other attributes*. The absence of even one required attribute in a test item will lead to an incorrect response to that test item. Simply put, the probability of correct response depends on the mastery of all the required attributes in a test item. For instance, an accurate response on a mathematical test item—[3 + 3.5 – 6 = 0.5]—would require mastery of both attributes, or, addition and subtraction.

Compensatory model. Compensatory models are sometimes referred to as disjunctive models, meaning that absence of mastery of an attribute can be compensated for by mastery of another or other attributes. This compensation suggests that the probability of correct response increases with the mastery of each of the attributes required by the test item. In other words, there is a cumulative increase in the likelihood of a correct response to a test item across the attributes. To clarify, let's present an arbitrary item—[4 x 3 + 6 = 18]—to see how one attribute can compensate for another to accurately respond to this item. Ostensibly, a student must have mastered both addition and multiplication attributes. Nevertheless, mastering only one of the required attributes could increase the probability of an accurate response: the knowledge of addition could compensate for the lack of knowledge of multiplication. To a student with only mastery of addition, this item may appear as [(4 + 4 +4) + 6 = 18].

Now that the two main categories of CDM have been explained, we can move on to the six subcategories, which will be further elaborated upon in the next six sections. The first three sections elaborate on the non-compensatory models and the following three are devoted to the three compensatory models.

Early research on CDM consisted of models that varied in approach and statistical analysis, while ended with similar results. The difference between these was in details of analyzes, e.g. estimating parameters at item level or attribute level. Although these models are similar in their estimation, the difference in approach made their comparison difficult and as a result, recent research has looked to integrating these models into a more general framework, the log-linear model (Rupp, Templin, & Henson, 2010).

Deterministic Inputs Noisy and-Gate Model (DINA)

DINA (Haertel, 1989; Junker, 1999; Junker & Sijtsma, 2001) is known for its parsimony, interpretability, and its easy fit to the data. Aside from other model parameters, there are two parameters common to the model, mainly guessing parameters g_j and slipping parameters s_j (de la Torre & Douglas, 2004). DINA model involves the three basic parameters of ξ_{ic}, g_i and s_i. *P*arameter ξ_{ic} represents the *deterministic input* for item *i* for an examinee in latent class *c*, indicating the examinee's array of attributes. Parameter g_i represents the probability of a correct response to item *i* for an examinee who has not mastered the required attributes for that item, that is, a "guessing" parameter. Parameter s_i represents the probability of an incorrect response to an item for an examinee who has mastered all the required attributes for item *i* i.e. *slipping* parameter. The deterministic parameter, guessing parameter, and slipping parameter for DINA model are defined as follows:

$$\xi_{\mathrm{ic}} = \prod_{a=1}^{A} \alpha_{ca}^{q_{ia}} \tag{4}$$

$$g_i = P(X_{ic} = 1 | \xi_{\mathrm{ic}} = 0) \tag{5}$$

$$s_i = P(X_{ic} = 0 | \xi_{\mathrm{ic}} = 1) \tag{6}$$

When an attribute is not measured by the item, the attribute ($\alpha_{ca}^{q_{ia}}$) becomes irrelevant, the q_{ia} expression will turn into zero and consequently the expression will equal to one. The multiplicative feature of this model indicates that an accurate response to a test item requires the mastery of all required attributes. Consequently, the probability of correct response to a test item by an examinee in latent class *c* is shown as:

$$\pi_{ic} = P(X_{ic} = 1 | \xi_{\mathrm{ic}}) = (1 - s_i)^{\xi_{\mathrm{ic}}} g_i^{1-\xi_{\mathrm{ic}}} \tag{7}$$

π_{ic} is the probability of correct response given the latent class membership.

Noisy-Input, Deterministic-and-Gate (NIDA)

The second non-compensatory model to review was introduced by Maris (1999): it aims at extending the DINA's inability to analyze items at the attribute level. Although still a non-compensatory model the probability is measured at the attribute level rather than the item level. Similar to DINA, NIDA has three main elements of slipping (s_a), guessing (g_a), and *deterministic input* for item *i* (ξ_{cia}). As the subscript of these elements *(a)* indicates the probabilities are at the skill levels rather than item level.

NIDA has local independence at the attribute level, meaning that measured attributes in an item are independent of one another. In other words knowledge of one attribute does not affect the knowledge of another. This property justifies the multiplicative property of the NIDA as follows:

$$s_a = P(\zeta_{\mathrm{cia}} = 0|\mathrm{a}_{\mathrm{ca}} = 1) \tag{8}$$

$$g_a = P(\zeta_{\mathrm{cia}} = 1|\mathrm{a}_{\mathrm{ca}} = 0) \tag{9}$$

$$\pi_{ic} = P(X_{ic} = 1|\alpha_c) = \prod_{a=1}^{A}\left[(1 - s_a)^{a_{ca}} g_a^{1-a_{ca}}\right]^{q_{ic}} \tag{10}$$

Parameter ζ_{cia} represents an examinee in latent class *c* either applying skill *α* correctly (ζ_{cia}*=1)* or incorrectly (ζ_{cia}*=0).* Parameter q_{ic} parameter can take on the value of 0 or 1 depending on whether the attribute is assigned to the item, making the attribute relevant or irrelevant to that item. The probability of correct response to item *i* for an examinee in latent class *c* is indicated by π_{ic}.

Non-Compensatory Reparameterized Unified Model (NC-RUM)

The last non-compensatory model to consider is NC-RUM, also known as the *fusion model* (Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007). NC-RUM is a slightly more complicated model in comparison to the previous two. It involves four probabilities, the probability of correctly applying a mastered skill required for item i (π_{ia}), the probability of correctly applying a non-mastered skill required for item *i* (r_{ia}), and the two main probability elements of the model, r_{ia}^* and π_i^*,that will be defined shortly. Rupp, Templin, and Henson (2010) related the first two probabilities to the guessing and slipping probabilities of DINA model. Their definition of these probabilities is as follows:

$$\pi_{ia} = P(\zeta_{\mathrm{cia}} = 1|\mathrm{a}_{\mathrm{ca}} = 1) = (1 - s_{ia}) \tag{11}$$

$$r_{ia} = P(\zeta_{\mathrm{cia}} = 1|\mathrm{a}_{\mathrm{ca}} = 0) = g_{ia} \tag{12}$$

Parameter π_{ia} represents the probability of correctly applying the attributes measured by an item for an examinee in latent class *c* given the mastery of attributes involved in that item. Parameter r_{ia} represents the probability of correct response to item *i* given the non-mastery of attributes involved in that item similar to guessing factor in the previous models. The other two main parameters for NC-RUM are defined as follows: parameter π_i^* represents the probability of answering item *i* correctly given the mastery of all attributes measured by item *i*, and parameter $\mathrm{r}_{\mathrm{ia}}^*$ represents the proportion of guessing over π_{ia} and are defined as:

$$r_{ia}^* = \frac{r_{ia}}{\pi_{ia}} \tag{13}$$

$$\pi_i^* = \prod_{a=1}^{A} \pi_{ia}^{q_{ia}} \tag{14}$$

Putting all the involved parameters together creates the unified model known as NC-RUM:

$$\pi_{ic} = P(X_{ic} = 1|\alpha_c) = \pi_i^* \prod_{a=1}^{A} r_{ia}^{*(1-\alpha_{ca})q_{ia}} \tag{15}$$

NC-RUM was the last non-compensatory model to consider in this literature review. The next three sections will cover the compensatory models starting with the simplest one in that subcategory.

Deterministic Input, Noisy-or-Gate (DINO) Model

The first compensatory model in CDM families to review is DINO, developed by Templin and Henson (2006). Similar to DINA, is fairly simple in its structure. It involves slipping (s_a) and guessing (g_a) parameters at the item level while utilizing latent response variable (ω_{ic}) to allow for one attribute to compensate for the other:

$$\mathrm{g_i} = \mathrm{P}(\mathrm{X_{ic}} = 1|\omega_{\mathrm{ic}} = 0) \tag{16}$$

$$\mathrm{g_i} = \mathrm{P}(\mathrm{X_{ic}} = 1|\omega_{\mathrm{ic}} = 0) \tag{17}$$

$$\mathrm{s_i} = \mathrm{P}(\mathrm{X_{ic}} = 0|\omega_{\mathrm{ic}} = 1) \tag{18}$$

$$\omega_{\mathrm{ic}} = 1 - \prod_{\mathrm{a}=1}^{\mathrm{A}} (1 - \mathrm{a_{ca}})^{\mathrm{q_{ia}}} \tag{19}$$

If an attribute was mastered ($i.e., a_{ca} = 1$), the latent response variable (ω_{ic}) will be equal to 1, increasing the probability of answering the item correctly. On the other hand, if the attribute was not mastered ($\mathrm{i.e.,a_{ca}} = 0$) the latent response variable (ω_{ic}) will be 0, decreasing the probability of answering the item correctly. Finally π_{ic} represents the probabilities of an accurate response for item *i* in latent class *c*:

$$\pi_{\mathrm{ic}} = \mathrm{P}(\mathrm{X_{ic}} = 1|\omega_{\mathrm{ic}}) = \ (1 - \mathrm{s_i})^{\omega_{\mathrm{ic}}} \mathrm{g_i^{1-\omega_{ic}}} \tag{20}$$

Noisy Input, Deterministic-or-Gate (NIDO)

The second compensatory model in cognitive diagnostic family to review is known as NIDO (Templin & Henson 2006). Its structure and composition is similar to NIDA: Its parameters are at attribute level rather than item level (Maris, 1999). The following is the expression for NIDO:

$$\pi_{\mathrm{ic}} = \mathrm{P}(\mathrm{X}_{\mathrm{ic}} = 1|\alpha_{\mathrm{ic}}) = \frac{\exp(\sum_{\mathrm{a}=1}^{\mathrm{a}}(\lambda_{.,0,(\mathrm{a})} + \lambda_{.,1,(\mathrm{a})}\alpha_{\mathrm{c}\alpha}\mathrm{q}_{\mathrm{i}\alpha})}{1 + \exp(\sum_{\mathrm{a}=1}^{\mathrm{a}}(\lambda_{.,0,(\mathrm{a})} + \lambda_{.,1,(\mathrm{a})}\alpha_{\mathrm{c}\alpha}\mathrm{q}_{\mathrm{i}\alpha})} \qquad (21)$$

In formula above, $\lambda_{.,0,(a)}$ represents the items intercepts, $\lambda_{.,1,(a)}$ represents the slope with an equality constraint across the items, π_{ic} represents the correct response to item *i* for examinees in latent class *c*, $q_{i\alpha}$ represents the Q-Matrix indicator, whether an attribute is measured by item *i* , and $\alpha_{\mathrm{c}\alpha}$ represents the attribute mastery for an examinee in latent class c.

It is important to note that this model can be subsumed under another family of models known as log-linear models, which will be discussed in more detail subsequently. The log-linear format is similar to the log-linear IRT expression. Although NIDO has the advantage of compensating for skills and formatting the model in the more familiar log-linear format, it has the disadvantage of constraining the intercept and slopes at the item level. To address this problem the next model was introduced, which will be the focus of this dissertation.

Compensatory Reparameterized Unified Model (CRUM)

CRUM is one of the most flexible compensatory models in CDM families (Rupp, Templin, & Henson, 2010). This model is the focus of this dissertation for its log-linear format, similarity to IRT models, and flexibility. Aside from its compensatory nature, it allows for items with the same Q-Matrix entries to have different parameters: Item intercepts ($\lambda_{i,0,(a)}$) and item slopes ($\lambda_{i,1,(a)}$*)* varies across the items, shown by *i* subscript. Unlike NIDO, slopes are defined at the attribute level, hence the kernel and probability is defined as:

$$\pi_{ic} = P(X_{ic} = 1|\alpha_{ic}) = \frac{\exp(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{c\alpha} q_{i\alpha})}{1 + \exp(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{c\alpha} q_{i\alpha})} \qquad (22)$$

π_{ic} represents the probability of an examinee in latent class *c* responding accurately to item *i*, parameter X_{ic} represents the observed response for item *i* for examinee in latent class *c*, parameter $q_{i\alpha}$ represents whether attribute α is represented in item *i*, parameter $\lambda_{i,0}$ represents the intercept for item *i*, indicated by 0 subscript, and parameter $\lambda_{i,1,(a)}$ represents the slope for item *i*, represented by subscript 1, at the attribute, *α,* level.

The focus of this study was on CRUM, based on its simplicity, flexibility, and accuracy. Recent studies on log-linear models have suggested a promising results from models with intercepts and main effects, rather than two and three way interaction effects (e.g., Choi, Templin, & Cohen, 2010; Kunina-Habenicht, 2010).Table 2.1 shows a summary of the six cognitive models in the CDM family. It includes a brief summary of each model, advantages, disadvantages, and their equations.

Estimation approach

Before continuing the literature review of CDM, it is important to discuss the two primary parameter estimation methodologies that have been used for these models: Markov Chain Monte Carlo (MCMC) and the estimation maximization (EM) algorithm. Patz and Junker (1999) argued that MCMC procedures are appropriate for log-linear models and complex format. They have argued that EM algorithm does not allow for incorporation of uncertainty (standard errors) into the item parameter estimation. Bayesian models on the other hand treat parameters as random variables and provides information on their posterior distribution—conditional distribution of the parameters, given the observed data.

More recently, however, researchers have utilized estimation maximization (EM) algorithms for these models. MCMC uses Bayesian techniques; EM uses marginal maximum likelihood (MML). Regardless of the algorithm used, both MCMC and EM require convergence checks. MCMC checks for convergence of a posterior distribution. EM checks for convergence within some specified tolerance.

Table 2.1

Summary of Cognitive Diagnostic Models

Model	Advantage	Disadvantage	Equation
DINA	Allows slip at item level	Correct only with mastery of all	$\pi_{ic} = P(X_{ic} = 1 \mid \xi_{\mathrm{ic}}) = (1 - s_i)^{\xi_{\mathrm{ic}}} g_i^{1-\xi_{\mathrm{ic}}}$
	Allows guessing at item level *deterministic input* (ξ_{ic})	Hence, does not distinguish which skills are lacked No matter what skills one lacks probability of missed items are equal Does not measure the skills difficulty	
NIDA	Measures probability of skills at an item	Its model fit depends on low probability of guessing and slipping factors.	$\pi_{ic} = P(X_{ic} = 1 \mid \alpha_c) = \prod_{a=1}^{A} \left[(1 - s_a)^{a_{ca}} g_a^{1-a_{ca}}\right]^{q_{ic}}$
	Allows slipping at skill level(s_a) Allows guessing at skill level((g_a) Local independence at skill level	Non-compensatory model	
Full NC-RUM	Allows slipping at skill level	Non-compensatory model	$\pi_{ic} = P(X_{ic} = 1 \mid \omega_{\mathrm{ic}}) = \ (1 - s_i)^{\omega_{\mathrm{ic}}} g_i^{1-\omega_{\mathrm{ic}}}$
	Allows guessing at skill level Incorporates the mis-designation of Q-Matrix	Fit of the model depends on high easiness parameter of	
DINO	Guessing parameter	Items with same Q-Matrix entries are equivalently parameterized $\lambda_{.,0,(a)}$	$\pi_{ic} = P(X_{ic} = 1 \mid \alpha_{\mathrm{ic}}) = \frac{\exp(\sum_{a=1}^{a}(\lambda_{.,0,(a)} + \lambda_{.,1,(a)}\alpha_{c\alpha})q_{i\alpha})}{1 + \exp(\sum_{a=1}^{a}(\lambda_{.,0,(a)} + \lambda_{.,1,(a)}\alpha_{c\alpha})q_{i\alpha})}$
	Slipping parameter Compensatory model Latent response variable (ω_{ic})		
C-RUM	Compensatory model	Intercept is defined at the item level $\lambda_{i,0}$	$\pi_{ic} = P(X_{ic} = 1 \mid \alpha_{\mathrm{ic}}) = \frac{\exp(\sum_{a=1}^{a} \lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)}\alpha_{c\alpha}q_{i\alpha})}{1 + \exp(\sum_{a=1}^{a} \lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)}\alpha_{c\alpha}q_{i\alpha})}$
	The slope is defined at the attribute level $\lambda_{i,1,(a)}$ The intercept can act as the guessing factor $\lambda_{i,0}$ The slope can act as the slipping factor $\lambda_{i,1,(a)}$		

Since information obtained from MCMC estimation is more comprehensive—MCMC estimation provide a full posterior distribution—than those obtained from an EM algorithm—point estimation and its standard error—the evaluation of convergence is more difficult (Roussos, Templin, & Henson, 2007). They further argued that, both approaches to estimations have been evaluated in simulation studies for efficacy in recovering item parameters in the cognitive models for which they were employed. Their convergence issues have been blamed on poorly specified Q-Matrices or few attribute representation (Templin & Henson, 2006).

Parameter estimation using both MCMC and EM have been studied with success; some of these existing studies using MCMC includes, Jang (2009) parameter estimation of a language test, de la Torre and Douglas (2004), and Henson, Templin, and Douglas' (2007) simulation study of parameter estimation. Some exiting studies using EM algorithm with success were Choi, Templin, and Cohen (2010) and Kunina-Habenicht's (2010) success in parameter estimation.

Although, MCMC creates more detailed output, the EM algorithm has many advantages and hence is more popular to use. Unlike MCMC, EM requires less time for the item parameters to converge. It uses a simpler algorithm. Researchers are also more familiar with its application and so it has been extensively applied to models with log-linear formats, mostly in IRT. In consideration of the reasons just mentioned, this study used EM algorithm for parameter estimation of CRUM.

Log-Linear Representation of CRUM

More recently researchers are inclined to organize the six core models as special cases of a more general model: to make them more compatible to the IRT formulations. For this purpose, von Davier's (2005) generalized diagnostic model (GDM) and then later Rupp, Templin, and

Henson's (2010) log-linear cognitive diagnosis model (LCDM) was introduced as the general format for the six core models of CDMs.

Their similarities to IRT models make them easier to understand and estimate using the traditional estimation techniques (such as MML with EM). Furthermore, a general format and approach helps with the estimation software, one software can estimate all six models.

One of the central ideas behind IRT is the estimation of person ability and item difficulty. Likewise, a CDM expressed in a log-linear format has item difficulty and person ability. CDM's item difficulty is defined as an item's intercept, the probability of correct response given non-mastery of all skill vector, indicated by the pre-assigned Q-Matrix, and CDM's person ability is defined as latent class profile mastery.

Among other cognitive models with a log-linear approach, the CRUM has consistently been demonstrated to recover its parameters more accurately under various simulation designs studies. Choi, Templin, and Cohen's (2010) study of several log-linear models suggested a superiority of CRUM's item recovery and misfit estimation over DINA, DINO, and LCDM. Simulation studies done by Kunina-Habenicht, Rupp, and Wilhelm (2012) and Choi, Templin, and Cohen (2010) have shown LCDM models tend to only recover their main effect parameters accurately. This suggests that interaction effect with combination to main effect can add complexity that can increase the estimation time of the LCDM models.

CRUM in Log-Linear Formulation

The CRUM in log-linear format is based on von Davier's (2005) general diagnostic model (GDM). Similar to other log-linear models, the CRUM consists of two main sub-formulas of item probability and priori classification. Item probability is also referred to as the measurement sub-formula and is as follows:

$$\pi_{ic} = P(X_{ic} = 1|\alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_i^T h(\alpha_c, q_i))}{1 + \exp(\lambda_{i,0} + \lambda_i^T h(\alpha_c, q_i))} \tag{24}$$

Let π_{ic} represent the probability of a correct response to item *i* for an individual with an attribute profile of α_c, let $\lambda_{i,0}$ represent the intercept of item *i*, and let λ_i^T represent the vector of slope for the main effect parameters for each item *i*. Similar to other CDMs, the CRUM has all three previously described essential structures: (1) attributes, (2) Q-Matrix, and (3) latent class attribute mastery profile.

The à priori classification sub-formula is also referred to as a structural model of log-linear models. It indicates the prior probability of class membership with only main effects in mind. The log-linear parameterization expression used to calculate the proportion of examinees in a particular latent class as follows:

$$\mu_c = \sum_{a=1}^{A} \gamma_{1,(a)} \alpha_{ca} \tag{25}$$

$$\upsilon_c = \frac{\exp(\mu_c)}{\sum_{c=1}^{c} \exp(\mu_c)} \tag{26}$$

In both formulas, μ_c represents the above expression that will be used to determine the proportion of the class membership for the population. The $\gamma_{1,(a)}$ represents the coefficient for the main effect associated with the attribute a presented in that particular class.

Putting the structural sub-formula with the measurement sub-formula of the log-linear model results in the final expression for calculating the probability in CRUM as follows:

$$(X_r = x_r) = \sum_{c=1}^{C} \vartheta_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1 - \pi_{ic})^{1 - x_{ir}} \tag{27}$$

Model Clarification a Practical Example

To clarify the mechanics of CRUM in log-linear format, a simple hypothetical mathematics exam has been created (van Davier, 2005). A hypothetical test with three items is shown in Table 2.2. This hypothetical test consists of two attributes, addition and subtraction. As Table 2.2 indicates, item one requires both addition and subtraction attributes for an accurate response, while item two requires only addition attribute for an accurate response.

Table 2.2

Three diagnostic hypothetical mathematics items

Item	Item	Addition	Subtraction
1	3+4-2	1	1
2	3+5	1	0
3	3-5	0	1

$$\boldsymbol{\pi_{ic} = P(X_{ic} = 1|\alpha_{ic}) = \frac{\exp(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{c\alpha} q_{i\alpha})}{1 + \exp(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{c\alpha} q_{i\alpha})}} \quad (28)$$

The above formula is used to calculate the probability of each item in CRUM in LCDM framework. The probability of correct response for item two, that measures the addition attribute only, for an examinee who has mastered the addition attribute involves two parameters of the intercept ($\lambda_{1,0}$ = -1.5) and the main effect ($\lambda_{1,1,(1)} = 2$) is as follows:

$$\pi_{1c} = P(X_{1c} = 1|\alpha_{1c} = 1) = \frac{\exp(-1.5\ + 2)}{1 + \exp(-1.5\ + 2)} = 0.62 \quad (29)$$

Using the same equation, the probability of correct response for item two for an examinee who has not mastered addition attribute is as follows:

$$\pi_{1c} = P(X_{1c} = 1|\alpha_{1c} = 0) = \frac{\exp(-1.5\ )}{1 + \exp(-1.5)} = 0.182 \quad (30)$$

The above calculation shows the missing of addition main effect. Using the same equation, the probability of correct response for item one for an examinee who has mastered both addition and subtraction is as follows:

$$\pi_{1c} = \mathrm{P}(\mathrm{X}_{1c} = 1|\alpha_{1c} = 1) = \frac{\exp(-1.5\ + 2 + 2)}{1 + \exp(-1.5\ + 2 + 2)} = 0.92 \quad (31)$$

Overview of Fit Indices

Another important issue that needs the researcher's attention is CDM's sensitivity to Q-Matrix misspecification. To evaluate accuracy of CDM in practice model's sensitivity to Q-Matrix misspecification is fundamental. Relative fit indices can be used to evaluate the fit of a model in comparison to the true model. Some of the existing studies regarding model fit will be reviewed in the following sections.

Fit indices, such as Akaike information criterion (AIC), Bayesian information criterion (BIC) are referred to as relative fit indices. They are mainly utilized for model comparison purposes. Their purpose is determining the best fitting model considering the most parsimonious model in mind (Kline, 2011). These particular indices are based on information theory and are often referred to as "penalized" model selection standards based on their formula involving a parameter and/or sample size term (Harrell, 2009). AIC is used to compare models that are fitted to the same data and it favors the model with the smallest value (Kline, 2005: Macready & Dayton, 1977).

These model comparison statistical analyses were first developed by Akaike (1973) for time series analysis. He stated that an hypothesis testing technique is not sufficient for statistical model identification. He utilized maximum likelihood estimation procedure to formulate a new information criterion (AIC) estimate, which was designed for the purpose of statistical model identification. Akaike presented an information criterion formulated as follows:

$$\text{AIC} = (-2)\log(\text{likelihood}) + 2(P) \tag{32}$$

In this equation "P" represents the number of parameters in the model. AIC is defined as an estimate of minus two times the log likelihood of the model whose parameters are determined by the method of maximum likelihood (Akaike, 1978). The value of this formula will increase as the number of parameters increases (Weakliem, 2004; Kuha, 2004). To select the best model, one considers the smallest values of each index indicate the best fitting model out of the models being proposed.

Akaike's (1978) stated that the introduction of AIC reveals the importance and practicality of modeling in statistics, but in spite of his model's widespread use, statisticians have expressed concerns regarding the application of this model. The creation of BIC was based on concerns regarding the application of AIC because of the frequent appearance of improper solutions, or occasions in which the parameters of the model could not have been estimated by the classical maximum likelihood method (Akaike, 1987).

To address this issue he later introduced a Bayesian interpretation of the AIC procedure. The Bayesian approach allows for incorporation of an equal prior probability distribution through the analysis of the likelihood function:

$$\text{BIC} = -2^*(\text{log likelihood}) + p^*\ln(N)\ (2) \tag{33}$$

In this equation N represents the sample size which was used to estimate the maximum likelihood of the function. While these two approaches were popular, the increase in the sample size resulted in favoring models with larger numbers of parameters/sample sizes. To address this issue an alternative information-based criteria was developed under the name of sample-size-adjusted BIC (Hurvich & Tsai, 1989). The sample size adjusted BIC is defined by Sclove (1987), where he substitutes the sample size n in the BIC formula with n* to eliminate the bias in the previous formula:

$$n^{*} = (n + 2) / 24 \qquad (34)$$

CDM Applications and Issues

To date, although there have been several successful applications of CDMs in the educational setting, the number of applications has remained fairly small and mostly have focused on mathematics. Some existing retrofitting analyses using cognitive models on mathematics assessments include studies by: Buck, Tatsuoka, and Kostin (1997); Galeshi & Skaggs (2010); Tatsuoka, Corter, and Tatsuoka (2004).

Buck, Tatsuoka, and Kostin's (1997) study indicated that Japanese a test of second language reading comprehension could be analyzed and successfully classify 91% of the examinees into their appropriate latent classes. Galeshi & Skaggs's (2010) study indicated that most items in block one of TIMSS (2007) mathematic could be successfully retrofitted with NCRUM, although, attributes with low representations were not able to be assessed accurately. Tatsuoka, Corter, and Tatsuoka's (2004) study of 20 countries indicated that Rule-Space methodology could successfully measure attribute mastery and compare the student's performance.

Retrofitting is fitting an existing dataset to CDM's statistical methodology to obtain informative inferences, but it has its criticisms. Although, retrofitting has been successful in many areas of assessments, its application has caused some concerns among researchers. Gierl, cui, and Zhou (2008) claimed that retrofitting CDM to most existing tests is likely to yield inaccurate diagnostic classification and despite its little success, it does not underestimate the great potential of problems with such an approach. Additionally, they have stated that retrofitting is fitting a new technology to an older apparatus. Since, existing tests have been designed based on older psychometric models with no regards to skill mastery and CDM's methodologies, retrofitting them could create great misclassification of student's mastery.

Results gained from retrofitting studies are subject and population specific, it is impossible to generalize them. Simulation studies on the other hand can provide researchers and practitioners with guidelines, which can be generalized to various populations. Considering these existing concerns regarding retrofitting, scarcity of CDM-designed tests, and a lack of great knowledge regarding these models' behavior, research on CDMs has so far been limited to a few simulation studies with a methodological focus (e.g. de la Torre & Douglas, 2004; Henson, Templin, & Willse, 2009). Their analysis of simulated data indicated that both EM and MCMC algorithm can successfully recover item parameters.

What Has Been Done

Despite the great recent interest in cognitive models, the research regarding CDMs is limited and random in model and topic of study. To date, there have not been any studies focusing specifically on CRUM. The closest study would be Choi, Templin, and Cohen's (2010) research that compared several log-linear models. Their study had focused on four various sample sizes, four attributes, and 25 items.

Two existing researches on evaluating fit under mis-specified Q-Matrix were Choi, Templin, and Cohen (2010) and Kunina-Habenicht, Rupp, and Wilhelm's (2012) study of CDM in log-linear format. Both of these studies utilized AIC and BIC indices to evaluate log-linear diagnostic models' relative fit with two types of Q-Matrix misspecifications and showed promising results. Choi, Templin, and Cohen (2010) reported that sample sizes of greater than 200 can effectively be evaluated for fit, Kunina-Habenicht, Rupp, and Wilhelm (2012) argued that parameter recovery and model fit was successful for 1,000 and 10,000 sample sizes. There is still a lot to be learned about how these models behave in real life testing situations where the practitioners cannot create a Q-Matrix with certainty.

One other existing study on fit was by Rupp and Templin (2008) on the effect of Q-Matrix misspecification on DINA model. They analyzed one dataset of four attributes and 15 items under True and over specification of the Q-Matrix indicated that DINA did not perform well under such a misspecification of the Q-Matrix.

Statement of the Problem

Despite CDM's popularity in recent years, their performance and parameter recoveries under various research designs is unknown (Choi, Templin, & Cohen, 2010; Roussos, Templin, & Henson, 2007). To be able to apply CDMs in operational testing situations, it is essential to study how accurately they can estimate a model's parameters. The existing gap in literature regarding CDM's parameter recovery and classification accuracy, hinder its application in practice. This study aims at establishing a rules of thumb guideline with regard to the number of items, number of attributes, and number of observations needed for valid application of CDM's.

One important issue regarding CDM's application is the specification accuracy of attributes by which items on a diagnostic assessment is measured. The specification of these attributes to items is represented in the Q-matrix. Its misspecification can have a detrimental effect on the parameter and classification accuracy of the model.

The correct specification of the Q-Matrix is a fundamental part of CDM applications. There is very little research done in analyzing the effect of misspecification on CDM's parameter recovery and its ability to accurately classify examinees into their accurate latent classes. In practice, it is almost impossible to verify the accuracy of a Q-matrix. There is little research in analyzing relative fit indices performance under Q-Matrix misspecification and the needed minimum number of examines with regards to attributes and test length. There is a need in examining the sensitivity of relative fit indices under various research designs of test length, observation number, and number of attributes.

The aim for this study is to attempt in studying the existing critical gap in the educational measurement literature using a complex simulation study evaluating parameter recovery and classification accuracy under accurate and mispecified QMatrices.

Chapter 3

Manuscript 1: Item Parameter Estimation Accuracy with a Cognitive Diagnostic Model: CRUM

Abstract

Item analysis is the heart of item response theory (IRT). Cognitive diagnostic models (CDM) expand IRT's results by investigating the attributes required by each item. Using a simulation study, the compensatory reparameterized unified model (CRUM), parameterized as log-linear model, was studied for its parameter recovery under various research designs using MPlus. This simulation study used samples of 50, 100, 500, 1,000, 5,000, and 10,000 examinees to evaluate the model's ability to recover its parameter and classify the examinees into accurate latent class mastery profile. These simulated datasets were as follows: 7 items with 3 attributes, 15 items with 4 attributes, 31 items with 5 attributes, and 35 items with 3 attributes. These test length/number of attribute combinations were chosen to incorporate all possible combinations and attributes in the dataset. The results indicated a strong relationship between the sample size and both parameter recovery and classification accuracy. The results also showed a strong relationship between the test length and both parameter recovery and classification accuracy.

Cognitive Models versus Traditional Psychometric Models

The traditional practice in education has been to identify students as proficient or non-proficient in a subject matter (de Ayala, 2009; Embretson & Reise, 2000). On the other hand, teachers and parents' interest have mainly focused on identifying areas of weakness to help students succeed (Mislevy, 1995). Although traditional test scores are useful in helping teachers and administrators evaluate student's performance, they do not provide adequate information about student's proficiency (Embretson, 1991; Mislevy, Almond & Lukas, 2004).Cognitive diagnostic model's (CDM) ability to produce a detailed report on examinees' skills has gained a great interest among teachers, psychologists and policy holders. CDMs aim to potentially provide specific targeted information on the type of knowledge or skills the examinee has or lacks on the subject matter. One example would be an item that requires addition and subtraction. An examinee could have mastered addition but not mastered subtraction; CDM allows for identification of mastery of these two attributes for each examinee depending on their overall performance on the test items.

Although test scores are generally useful measures of students' ability and item difficulties, it fails to produce a measure of student's cognitive processes (Norris, Macnab, & Phillips, 2007). Other traditional test theories such as classical test theory (CTT) (de Ayala, 2009) also provide a summative analysis, an investigation on classification rather than creating a causal path from the direct observable behaviors (Harre, 1970). To understand achievement test scores, an explanation of student's performance is required. Cognitive models are suited for explaining the underlying detail of the student's particular performance. CDMs have the potential to provide valuable information from standardized-based assessments, such as student's attribute mastery and diagnostic feedback (Rupp & Templin, 2008). This implication is significant for complying with the mandated “No Child Left Behind” policy in the United States.

Rationale for the Study

CDMs are relatively new to the field of psychometrics, which means a great deal of research is needed to truly understand their results and performance. In recent years, research in improving and applying diagnostic models has been rapidly increasing mainly due to the nature of information that these approaches can provide (Rupp, Templin, & Henson, 2010). Little is known about their behavior under various research designs. Since these models are new, it is important to study their behavior under various test conditions. There is no specific research dedicated on studying the CRUM under various conditions, and little is known about these models’ behavior. The aim of this study is to address this critical gap by focusing on parameter recovery and classification accuracy of the CRUM parameterized as a log-linear model. A vast amount of research is needed to examine these models’ validity and reliability, specifically in educational and psychological settings.

Research has shown sample size, test length, and number of attributes can affect the parameter and classification recovery of cognitive models (e.g. Choi, Templin, & Cohen, 2010; Kunina-Habenicht, 2010; von Davier’s, 2005). The aim of this study is to investigate the effect of sample size, item length, and attributes complexity in recovering item parameters under log-linear framework. The two specific research questions are:

1. How accurately does the model recover its parameters under various test length/ Q-Matrix and sample size conditions when dichotomous data is generated from a CRUM?

2. How accurately does the model classify students into their true class of attribute mastery under various test length/ Q-Matrix and sample size conditions when dichotomous data is generated from a CRUM?

An Overview of Cognitive Diagnostic Models

A CDM's main purpose is to create detailed information regarding the students' cognitive strengths and weaknesses. CDM is a fairly new methodology, and only a few studies have investigated its performance under various research designs. Some of the existing studies of cognitive models are de la Torre and Douglas (2008) for the DINA model and Henson, Roussos, and Templin (2005) for the RUM model. These studies were limited to a few observation levels and models based on non-linear approaches. This study focuses on a log-linear approach to one of the core CDMs, the Compensatory Reparameterized Unified Model (Henson, Templin, & Willse, 2009), and it investigates this model's parameter recovery and classification accuracy under various conditions.

CDMs are psychometric models with the potential for providing rich information to improve teaching and learning (de la Torre & Douglas, 2008). However, wider applications of these models have been hampered by their novelty. As von Davier (2005) explained, the intention of testing analysis is to make valid inferences about individual differences on the measured subject matter. Most diagnostic classification models implement a discrete multivariate latent variable that signifies the absence or presence of multiple attributes.

CDMs are probabilistic diagnostic models for analysis of binary data. It enables researchers to create a link between examinee's binary response and their underlying attribute/skill or ability on the subject matter (Rupp, Templin, & Henson, 2010). CDM quantifies examinee's latent trait for the purpose of classification of these examinees into their latent class mastery. To date, researchers have suggested various models and approaches for measuring student's attributes. Regardless of approach, these models share some fundamental structures that will be discussed in the following section.

Log-linear CRUM

Common to all CDMs is a fundamental input known as the Q-Matrix. A Q-Matrix is a loading matrix connecting the attributes to their relevant items. Each Q-Matrix uses attributes for its elements. Attributes first and foremost are unobservable: they are constructs only estimable from observation and they are measured by item responses. Seen as how measurements are made indirectly, Q-Matrix construction requires subject experts, intense consideration, and careful attention to its specification, which would otherwise falsify the results (Tatsuoka, 1983).

Posterior distribution is a common output. Cognitive models classify students into their accurate latent class mastery to individualize feedback, to inform students of their non-mastered attributes, and to retain through practice their mastered attributes (Huff & Goodman, 2007). If a student had mastered proportions but does not quite understand graphs and geometry, CDM would provide this knowledge. The posterior distribution estimates the relative frequency of examinees across all possible latent classes, or profiles of mastery or non-mastery of each attribute.

The log-linear CRUM was originally based on Von Davier's (2005) generalized diagnostic model (GDM) and then was later furthered by Rupp, Templin, and Henson's (2010) log-linear cognitive diagnostic model (LCDM). The LCDM model (and its variations) relates existing cognitive models to the well-practiced technique of IRT by way of people's ability (θ) and the item difficulty (b) that correlates to CRUM's item intercept (β_{0i}) and latent class profile mastery (C).

The CRUM uses a logistic link function to cohere with linear predictors in any given regression model. Here, the independent variable expresses the combination of the latent classes and item parameters—the intercept and the main effect. Unlike CRUM, Henson, Templin, and Willse's (2009) LCDM incorporates both main and interaction effects. The general logistic link function is as follows:

$$logit = \log[\frac{P(X = x\ |\beta_i, q_i\gamma_i\alpha)}{P(X = 0\ |\beta_i, q_i\gamma_i\alpha)}] = \beta_{0i} + \gamma_{xi}^T h(q_i, \alpha)$$

Similar to IRT model, the logit pictured contains the independent predictors—$h(q_i, \alpha) = (h_1(q_i, \alpha, \dots \dots h_k(q_i, \alpha))$—and combines item specifications—q_i ∈ {0,1}—with the person's ability—$\theta = (a_1, \dots .. a_k)$. Lastly γ_i indicates slope for each attribute.

CRUM consists of two sub-formulas: probability sub-formula and measurement sub-formula. Probability sub-formula embeds the logit above in the equation to estimate the probability. (See the IRT format below.)

$$P(X_{ir} = 1|\beta_{ik}, q_{ik}, \alpha_r) = \frac{\exp(\beta_{0i} + \sum_{k=1}^{K} \beta_{1ik} q_{ik} \alpha_{rk})}{1 + \exp(\beta_{0i} + \sum_{k=1}^{K} \beta_{1ik} q_{ik} \alpha_{rk})}$$

In the formula, X_{ir} represents the observed response of individual *r* in the latent class *c* to item *i*; β_{0i} represents the intercept, considered as the overall item difficulty at the item level; β_{1ik} represents the attribute's slope at the item level as indicated by the subscripts *i* and *k*; and, finally, α_{rk} represents the latent class mastery specification.

To demonstrate, a latent class with attribute *k* mastery is represented by α_{ik}=1; otherwise, attribute *k* non-mastery is represented by α_{ik}=0, meaning, for an examinee in that latent class, there is a lower probability of accurate response. Moreover, an item with attribute *k* has the slope β_{1ik}=1; otherwise, an item with attribute *k* has the slope β_{1ik}=0, meaning, mastery of the attribute is irrelevant to that item.

Moving on, the second sub-formula is the measurement sub-formula. Of the two formulas pictured below, the mixing kernel μ_c uses the marginal probability of the latent classes to calculate the measurement sub-formula υ_c:

$$\mu_c = \sum_{a=1}^{A} \gamma_{1,(a)} \alpha_{ca}$$

$$\upsilon_c = \frac{\exp(\mu_c)}{\sum_{c=1}^{c} \exp(\mu_c)}$$

Seen above, $\gamma_{1,(a)}$ represents the coefficient of regression for the main effect associated with the attribute presented in that particular class, and υ_c represents the proportion of population belonging to that specific latent class over all possible latent classes.

Combining the structural sub-formula with measurement sub-formula of the log-linear model will result in the following CRUM:

$$P(X_r = x_r) = \sum_{c=1}^{C} \vartheta_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1 - \pi_{ic})^{1-x_{ir}}$$

Seen above, $\pi_{ic}^{x_{ir}}$ represents the probability of an accurate response given the observed response to item *i*, and x_{ir} represents examinee *r*'s observed response.

Cognitive diagnostic approach. Cognitive diagnostic parameter estimation is mostly done with two approaches of Markov-chain Monte Carlo (MCMC) in a Bayesian framework and the expectation maximization (EM) algorithm ((Roussos, Templin, & Henson, 2007). Although, Patz and Junker (1999) have argued that Monte Carlo procedures are appropriate for models with log-linear design and complex format, more recently researchers are inclined to use EM algorithm for its familiarity and well known approach of marginal maximum likelihood (MML). Estimation approaches using both EM and MCMC for cognitive models have been evaluated with success with simulations studies for effectiveness of recovery of item parameters (Roussos, Templin, & Henson, 2007).

CRUM is a stochastic model—the probabilistic counterpart to a deterministic model—mainly used for dichotomous items in which the probability of observing a specific response vector is estimated as a function of a set of *C* latent classes with the assumption of independence within the latent classes (Templin, Henson, Rupp, Jang, & Ahmed, 2009).

Research on Parameter Recovery and Classification

There have been a few studies addressing parameter recovery and classification issues on various cognitive models. However, there have not been many studies focusing specifically on CDM in log-linear format. Choi, Templin, and Cohen's (2010) simulation study of parameter recovery of four log-linear models has shown that the CRUM was the most effective in parameter and classification recovery of the models studied even when the data were generated under the full log-linear model.

A study by Kunina-Habenicht, Rupp, and Wilhelm (2012) was performed on simulated datasets of 1,000 and 10,000 examinees with 25 and 50 items and three and five attributes. The results suggested that datasets with five attributes and smaller sample sizes did not recover its parameter as well as larger datasets with three attributes. But similar to the previous studies the main effect was estimated more accurately than the interaction effect, suggesting that using models with only the main effect would be more practical.

Another study in log-linear format is von Davier's (2005) simulation study of a dataset with 36 items, four attributes, and 2880 examinees with 40 replications. This study showed that the classification as well as item parameters were accurately recovered under the full log-linear specification. Templin, Henson, Rupp, Jang, and Ahmed's (2009) study on the CRUM using

Markov Chain Monte Carlo (MCMC) approach showed that CRUM can recover item parameters when the items are dichotomous and the sample size is 1,000 examinees or higher. They examined the behavior of the CRUM with 25 items, four attributes, and sample sizes of 500, 2,000, 5,000, and 10,000 with a maximum of two attributes per item using an MCMC algorithm and 25 replications.

Method

Although three to five attributes may be considered low in practice, it is considered a reasonable number of attributes in cognitive models research. Existing literatures in latent class analysis are limited to four to eight attributes based on the required computational intensity of these models (Hartz, 2002; Maris, 1995; Rupp & Templin, 2008; Templin & Henson, 2006). As the number of attributes increases linearly, the number of latent classes increases exponentially, leading to an upsurge in the number of iterations needed for the parameters to converge (Rupp & Templin, 2008). For example, a study conducted by Templin and Henson (2006) on the DINA model with 40 items, seven attributes, and 3,000 respondents, acquired four days to complete.

This study used 200 replications: Harwell, Stone, Hsu, and Krisci (1996) discussed the number of necessary replications for attaining reliable and stable results in simulation studies in an IRT framework with MCMC methodology and recommended a minimum of 25 replications for achieving reliable results. They further stated that the number of replications needed for reliably estimating effects is higher when the complexity of the effects increases (e.g., main, two-way, and three-way effects). The complexity of this study is limited to the intercept and main effect, hence 200 replications is more than what has been done or recommended in previous research for cognitive models (e.g. de la Torre & Douglas, 2008; Rupp & Templin, 2008).

Data were simulated using SAS version 9.2 and analyzed with Mplus version 5 (Muthen & Muthen, 2010). To simulate item response data, two main distributions are required: the person's ability and item difficulty.

Manipulated Factors

Sample size. The first independent variable was the number of respondents. To date, researchers have not been able to establish a guideline regarding the number of examinees needed for an accurate application of CDMs. Sample sizes were selected with the consideration of practical implementation. CDMs are created to enable practitioners gain formative information on their student's performance. Hence, it is practical to examine the behavior of this model under small sample size of 50 or large sample of 10,000 examinees, since these models are designed to provide feedbacks at school or district level for a possible ongoing benchmark testing purposes. Furthermore, research has shown a strong relationship between reliability of the fit and the number of examinees (e.g. Choi, Templin, & Henson, 2010; Kunina-Habenicht, Rupp, & Wilhelm, 2012).

Within many cognitive diagnostic models CRUM in log-liner format has not been studied extensively under various sample sizes. To establish a guideline a wide range of 50, 100, 500, 1,000, 5,000, and 10,000 sample sizes were examined under an accurate Q-Matrix. Existing studies on various cognitive models have 1,000 to 50,000 examinees and have focused on many cognitive models except CRUM in log-linear format (e.g. de la Torre, 2008; Kunina-Habenicht, Rupp, & Wilhelm, 2012). Extending the range of the sample sizes to 50 and 10,000 examinees can help researchers and practitioners to follow a rule of thumb for the CRUM in practical settings.

Number of attributes in consideration of test length. The second independent factor was the number of attributes in combination with test length. Three to five attributes are considered a reasonable number of attributes in cognitive model research and in practice. While, three attribute exams can be administered as a monthly instructional assessment purposes, five attribute exams can be administered as a benchmark progress report. It is necessary to remind the readers that the main purpose of implementing CDM assessment technique is to provide continuous feedback to students and instructors.

Existing research has shown the importance of attributes and test length in CDM's estimation (e.g. Lai, Gierl, & Cui, 2012; Rupp & Templin, 2008). Similarly, the effect of number of attributes per item, or item complexity, can affect the parameter estimation. The numbers of items were selected in such a way that all possible attribute combinations were included in the dataset. It was important to the purpose of this research to include all possible combinations of the items. To examine the effect of test length on the parameter and the classification accuracy, dataset with 7 versus 35 items with 3 attributes were examined.

Table 3.1

Summary of Treatment Designs and Observation Design

Manipulated Parameter	Assigned value
Number of observations	N=50, 100, 500, 1,000, 5,000, & 10,000
Number of attributes	k= 3, 4, 5
Number of items/attributes	j= 7/3, 15/4, 31/5, 35/3
Number of replications	200
Number of simulation studies	6*4=24
Number of observation design	24*200=48,000

To examine the effect of item complexity, the intercept values were manipulated to vary between -1.5, the easiest item with one attribute per item, to -7.59, the hardest item with five attribute per item (Rupp, Templin, & Henson, 2010). To clarify item complexity consideration, a dataset with three attributes requires a minimum of seven items to include all possible attribute per item combinations, 2^3 -1 = 7 possible combinations, with one combination of no occurring skills (0,0,0). An item with one required attribute is considered to be an easier, less complex, item than the one with three attributes.

Table 3.2

The Q-Matrixes Used for Simulating Data

J=7, K=3		J=15, K=4		J=31, K=5		J=35, K=3	
Items	Q-Matrix	Items	Q-Matrix	Items	Q-Matrix	Items	Q-Matrix
1	0 0 1	1	0 0 0 1	1	0 0 0 0 1	1	0 0 1
2	0 1 0	2	0 0 1 0	2	0 0 0 1 0	2	0 1 0
3	0 1 1	3	0 0 1 1	3	0 0 0 1 1	3	0 1 1
4	1 0 0	4	0 1 0 0	4	0 0 1 0 0	4	1 0 0
5	1 0 1	5	0 1 0 1	5	0 0 1 0 1	5	1 0 1
6	1 1 0	6	0 1 1 0	6	0 0 1 1 0	6	1 1 0
7	1 1 1	7	0 1 1 1	7	0 0 1 1 1	7	1 1 1
		8	1 0 0 0	8	0 1 0 0 0	8	0 0 1
		9	1 0 0 1	9	0 1 0 0 1	9	0 1 0
		10	1 0 1 0	10	0 1 0 1 0	10	0 1 1
		11	1 0 1 1	11	0 1 0 1 1	11	1 0 0
		12	1 1 0 0	12	0 1 1 0 0	12	1 0 1
		13	1 1 0 1	13	0 1 1 0 1	13	1 1 0
		14	1 1 1 0	14	0 1 1 1 0	14	1 1 1
		15	1 1 1 1	15	0 1 1 1 1	15	0 0 1
				16	1 0 0 0 0	16	0 1 0
				17	1 0 0 0 1	17	0 1 1
				18	1 0 0 1 0	18	1 0 0
				19	1 0 0 1 1	19	1 0 1
				20	1 0 1 0 0	20	1 1 0
				21	1 0 1 0 1	21	1 1 1
				22	1 0 1 1 0	22	0 0 1
				23	1 0 1 1 1	23	0 1 0
				24	1 1 0 0 0	24	0 1 1
				25	1 1 0 0 1	25	1 0 0
				26	1 1 0 1 0	26	1 0 1
				27	1 1 0 1 1	27	1 1 0
				28	1 1 1 0 0	28	1 1 1
				29	1 1 1 0 1	29	0 0 1
				30	1 1 1 1 0	30	0 1 0
				31	1 1 1 1 1	31	0 1 1
						32	1 0 0
						33	1 0 1
						34	1 1 0
						35	1 1 1

Table 3.1 shows the summary of the research designs as follows: three attributes with seven and 35 items, four attributes with 15 items, and five attributes with 31 items. Each dataset was manipulated for six sample sizes with 200 replications, resulting in 24 * 200 = 48,000 datasets to investigate.

Although there are some existing studies on the accuracy of parameter recovery in log-linear models, this study differs from the previous ones by its combination of model, estimation technique, sample sizes, as well as item/attribute combinations. Also, unlike most existing literature, it incorporates all possible combinations of attributes per items, as shown in Table 3.2.

To date, there has not been any existing study dedicated to extensive investigation of CRUM in log-linear format. The study included the manipulation of sample sizes with regards to attributes and item difficulty. The item lengths in consideration of attribute were as follows: three attributes with seven and 35 items, four attributes with 15 items, and five attributes with 31 items. To ensure the accuracy of estimation an extra dataset with 35 items was included, considering the short test length of seven items with three attributes.

The similarity between CRUM in log-linear format can incorporate the two IRT parameters as person ability and item difficulty. Person ability in a diagnostic model is indicated by latent class membership, while item difficulty is generally viewed as the value of the intercept.

Item Difficulty

Purposeful and uniform distribution for the item difficulty was employed for the generation of the simulated data. An overall/baseline item difficulty in log linear model is measured by the intercept of the item (λ_0), indicating the probability of responding to an item correctly given that the respondent have not mastered any of the required skills for that particular item. Hence, items with larger numbers of attributes will tend to have higher values for the intercept, indicating more difficult items, and items with small number of attributes will tend to have lower value for the intercept, indicating easier items.

Item difficulty was another aspect of these models. After purposely selecting the item difficulties, an equal weight was assigned to each attribute. Table 3.3 shows that all main effects, attributes, have the same value (λ_1=2). This means that all attributes were equally difficult/easy. Hence, the probability of correct response increases uniformly as the number of acquired attributes increases for an individual (Table 3.3).

Table 3.3

Main Effect and Intercept Parameters Used in Simulation Design

	Item difficulty(λ_0)	Main effect(λ_1)
Items with 1 attributes	-1.5	2
Items with 2 attributes	-2.5	2
Items with 3 attributes	-3.375	2
Items with 4 attributes	-5.063	2
Items with 5 attributes	-7.59	2

Person Ability

Another important aspect of this simulation study is deciding on the person ability distribution. Person ability in cognitive diagnostic model is the membership in a latent class within the model. Each latent class has specific latent attribute profile indicating mastery or non-mastery of each attribute assigned to the test. A uniform distribution of latent classes was selected to account for all possible real life situations, except for the last class that had to be adjusted for the remaining probability, an Mplus requirement (Rupp, Templin, & Henson, 2010).

Table 3.4 shows the person ability profile mastery and the proportion of the population that falls into that profile. To distribute the simulated proportion a purposeful uniform distribution was determined to ensure an equal proportion for all attribute profiles, υ_c.

Table 3.4

Class Membership Proportion Used in Simulation Design

Latent profile "C"	Class #	$\upsilon_{\mathbf{c}}$/Prob	$\boldsymbol{\mu}_{\mathbf{c}}$
(0,0,0)	1	0.25	1.00
(0,0,1)	2	0.083	1.00
(0,1,0)	3	0.083	1.00
(1,0,0)	4	0.083	1.00
(0,1,1)	5	0.083	1.00
(1,0,1)	6	0.083	1.00
(1,1,0)	7	0.083	1.00
(1,1,1)	8	0.25	0.00

Evaluation Technique

The two research questions focus on the CRUM's ability to recover item parameter and individual's classification. To evaluate the accuracy of item parameters recovery, point estimation, estimated bias, the standard error (SE), and the root mean square error (RMSE) were calculated. The bias estimation is the mean difference between the estimated value and the true value over 200 replications:

$$avg\ Bias = \frac{1}{200}\sum_{r=1}^{200}(\hat{\lambda}_r - \lambda)$$

RMSE is used to evaluate the total variance of estimation error between the observed item parameters and the true item parameters (Steiger, 1990). The measuring index is as follows, zero indicates no discrepancy between the model and the true value, in other words, suggesting a true fit of the model to the data. On the other hand, any value of RMSEA > 0.1 would indicate a poor fit (Oliveri & von Davier, 2011). The recovered values and their standard errors are averaged over the 200 replications; the RMSE of the parameter estimates is calculated as follows:

$$RMSE\left(\hat{\theta}\right) = \sqrt{\frac{1}{200}\sum_{t=1}^{200} E(\hat{\lambda}_{ti} - \lambda_{true})^2}$$

where *i* is the number of items, $\hat{\lambda}$ is the estimated parameter and λ_{true} is the true value for the parameter.

Evaluating Classification

Classification of individual examinees into a specific class and hence their skill mastery profile is the most valuable information gained from CDMs analysis and therefore the accuracy of a mastery profile is crucial in evaluating model parameter estimation. Mastery accuracy was studied at the group level to evaluate degree of classification accuracy. To evaluate group level accuracy, an average classification proportion bias of 200 replications was calculated as well as the correlation between the true classification proportion and estimated classification proportion.

The bias for the classification probability was calculated by evaluating the difference between the estimated classification probability, $\hat{\lambda}_r$, value and the true classification probability, λ_r as follows:

$$avg\ Bias = \frac{1}{200}\sum_{r=1}^{200}(\hat{\lambda}_r - \lambda_r)$$

A correlation between the true latent class membership probability and estimated latent class membership was also calculated and the results are included in the analysis.

Results

The results section is divided into two sections, each focuses on one research question. The first part of the result analysis the item parameter recovery and the second part analysis the results from classification of the examinees under various sample sizes and test length with combination of number of attributes. The key findings are included into three tables of intercept, main effect, and classification with bias, SE, and RMSE values included.

Before starting the analysis we note that parameter estimation results for 100 and 50 observations, were not shown here based on the inaccuracy of the estimation and the high percentage of extreme bias values except for dataset with three attributes and 35 items. This occurrence is as the result of low representation of the attributes among items as well as sparse cells for classification purposes. To clarify, datasets with 31items, five attributes, and 50 examinees did not have sufficient number of representations for all 32 latent classes.

One example of the unreliable results was the datasets with 50 observations, three attributes, and seven items, which resulted in only 30% of the item parameters, were recovered accurately with low SE. Another example of extreme bias estimation was datasets with four attributes and 15 items. About 35% of the parameters for 100 observations had extremely high estimation bias of greater than 20 points.

Parameter Recovery

The first research question was designed to investigate the accuracy of the model in recovering the item parameters under various test length/ Q-Matrix and sample sizes under CRUM specification. The results are consistent with existing literature regarding the accuracy of main effect and intercept parameters where study conditions overlap (Kunina-Habenicht, 2010; Choi, Templin, & Cohen, 2010). Table 3.5 compares the estimated intercept and Table 3.6 compares the estimated main effect parameters for datasets of 500, 1000, 5000, and 10,000 examinees. As mentioned above, the analyses for 100 and 50 examinees were not shown in the table based on frequency of inaccurate recovery for all datasets except for dataset with three attributes and 35 items.

In Table 3.5, the results show that datasets with 10,000 examinees had the most precise point estimation with low SEs. At first glance, a noticeable result is the improvement of parameter estimation with the increase of sample sizes. Secondly, as the number of items increases the accuracy of estimation improves as it is shown for datasets with seven and 35 items with three attributes (RMSE<0.1).

As the number of attributes increases to four and five per dataset, the item complexity/difficulty increases. Table 3.5 summarizes the intercept effect by item specification. The results indicated that the intercepts were estimated accurately for datasets of 10,000 examinees regardless of the data specifications. One important result to notice is the dataset with 35 items. Its results indicate that regardless of the number of examinees, it tends to recover item intercepts accurately.

Another important result to note is the negative relationship between item complexity and the parameter recovery; as the items become more difficult, item complexity increases, the accuracy of the item parameter recovery decreases significantly. This result is common for both main effect parameters (discussed below) as well as intercepts parameters. Items with less than three attributes tended to recover their parameters with low bias (bias < 0.08) except for very short dataset of seven items and three attributes. Moreover, items with four and five attributes have the highest estimation bias, requiring more than 5,000 observations for accurate estimation (RMSE <0.1).

To summarize the results, parameter recovery for items with three attributes require minimum of 10,000 observations, while datasets with 35 items recovered its intercept with observation as little as 500 examinees and less. The results from main effect parameter recovery are similar to intercept parameter recovery. Test length with seven items did not perform well for datasets with low observations of 500 examinees, while parameters were recovered well for datasets with high observation number of 5,000 and higher.

Table 3.5

A Comparison of Intercept Parameters

	7 Items with 3 attributes				35 Items with 3 attributes			
	N=10,000	N=5,000	N=1,000	N=500	N=10,000	N=5,000	N=1,000	N=500
	3 Items Required one attribute				15 item Required one attribute			
Bias λ0	0.02	0.04	0.02	0.08	0	0.002	0.01	0.016
SE	0.08	0.12	0.31	0.58	0.038	0.054	0.123	0.176
RMSE	0.053	0.09	0.23	0.49	0.0001	0.00018	0.0009	0.0024
	3Items required two attributes				15 items required two attributes			
Bias λ0	0.07	0.08	0.06	0.11	0.084	0.086	0.097	0.058
SE	0.11	0.16	0.42	0.84	0.053	0.075	0.168	0.244
RMSE	0.06	0.1	0.36	0.7	0.0003	0.00051	0.0019	0.0057
	1 Items required three attributes				5 items required three attributes			
Bias λ0	0.02	0.09	0.05	0.17	0.084	0.012	0.048	0.024
SE	0.14	0.21	0.52	0.92	0.053	0.101	0.231	0.327
RMSE	0.07	0.18	0.45	1.08	0.0011	0.0028	0.014	0.032
	Items=31 & Skills=5				Items=15 & Skills=4			
	N=10,000	N=5,000	N=1,000	N=500	N=10,000	N=5,000	N=1,000	N=500
	5 items Required one attribute				4 item Required 1 attribute			
Bias λ0	0	0	0	0.01	0	0	0.01	0.06
SE	0.04	0.06	0.14	0.21	0.07	0.09	0.21	0.37
RMSE	0.00033	0.0158	0.0034	0.012	0.0008	0.002	0.009	0.053
	9 items required two attributes				6 items required 2 attributes			
Bias λ0	0	0	0.03	0.02	0.01	0	0.01	0.08
SE	0.06	0.09	0.21	0.31	0.07	0.09	0.3	0.37
RMSE	0.00041	0.00088	0.0051	0.022	0.002	0.003	0.019	0.061
	10 items required three attributes				4 items required 3 attributes			
Bias λ0	0	0	0.04	0.07	0	0.01	0.07	0.21
SE	0.09	0.12	0.28	0.43	0.1	0.13	0.4	0.51
RMSE	0.00089	0.002	0.0111	0.029	0.0041	0.009	0.057	0.18
	5 items require four attributes				1 item required 4 attributes			
Bias λ0	0.01	0.02	0.06	0.18	0.01	0.03	0.12	0.25
SE	0.12	0.17	0.39	0.59	0.14	0.2	0.5	0.83
RMSE	0.00258	0.007	0.035	0.098	0.0179	0.046	0.24	0.45
	2 items require five attributes							
Bias λ0	0.04	0.07	0.13	0.04	-	-	-	-
SE	0.18	0.25	0.58	0.87	-	-	-	-
RMSE	0.0337	0.0577	0.353	0.5	-	-	-	-

Classification Estimation

The second research question aimed at investigating the accuracy of the classification of examinees into latent classes under various sample sizes as well as test length/ number of attributes. Classification of student’s attribute mastery is the most important outcome of these models and its results needs special attention. Similar to the previous results, the analysis for 100 and 50 observations were not included in the tables based on the inaccuracy of parameter estimations.

Table 3.6

A Comparison of Main Effect Parameters

	Data set with I = 35 & K = 3				Data set with I = 7 & K = 3			
	N=10,000	N=5,000	N=1,000	N=500	N=10,000	N=5,000	N=1,000	N=500
	15 item Required one attribute				3 Items Required one attribute			
Bias	0.0039	0.06	0.015	0.203	0.04	0.04	0.14	0.18
SE	0.0097	0.08	0.046	0.066	0.08	0.07	1.12	0.58
RMSE	4.50E-06	6.5E-06	0.0003	0.042	0.17	0.18	0.46	0.49
	15 items required two attributes				3Items required two attributes			
Bias	0.0041	0.06	0.015	0.204	0.19	0.24	0.29	0.30
SE	0.009	0.08	0.17	0.066	0.08	0.09	0.43	0.84
RMSE	4.80E-06	4.8E-05	0.0003	0.042	0.02	0.39	0.21	0.7
	5 items required three attributes				1 Items required three attributes			
Bias	0.0048	0.01	0.02	0.233	0.24	0.35	0.40	0.47
SE	0.0169	0.08	0.09	0.13	0.08	0.12	0.32	0.92
RMSE	3.90E-05	7.9E-05	0.0007	0.054	0.31	0.32	0.44	1.08
	N=10,000	N=5,000	N=1,000	N=500	N=10,000	N=5,000	N=1,000	N=500
	Data set with I = 15 & K= 4				Data set with I = 31 & K = 5			
	4 item Required one attribute				5 items Required one attribute			
Bias	0	0.01	0.03	0.09	0	0	0.01	0.201
SE	0.07	0.07	0.09	0.30	0.05	0.07	0.21	0.257
RMSE	0	0	0	0.05	0.02	0	0	0.073
	6 items required 2 attributes				9 items required two attributes			
Bias	0.06	0.09	0.03	0.10	0.01	0.01	0.04	0.02
SE	0.07	0.09	0.35	0.37	0.06	0.08	0.21	0.35
RMSE	0	0	0.02	0.06	0	0	0	0.02
	4 items required 3 attributes				10 items required three attributes			
Bias	0	0.02	0.238	0.22	0.01	0.03	0.07	0.10
SE	0.01	0.13	0.4	0.51	0.09	0.12	0.22	0.43
RMSE	0	0.01	0.059	0.18	0	0	0.05	0.10
	1 item required 4 attributes				5 items require four attributes			
Bias	0.01	0.104	0.252	0.26	0.06	0.06	0.19	0.28
SE	0.06	0.2	0.5	0.83	0.09	0.09	0.39	0.59
RMSE	0.02	0.056	0.31	0.45	0	0	0.06	0.10
-	-	-	-	-	2 item required 5 attributes			
-	-	-	-	-	0.10	0.10	0.23	0.32
-	-	-	-	-	0.16	0.16	0.58	0.80
-	-	-	-	-	0.02	0.02	0.36	0.30

Table 3.7 is a summary of classification accuracy at the group level. It compares the correlation between estimated classification probability and true classification probability at the group level as well as probability bias for classification. The true proportion of individuals in a specific class was correlated with estimated proportion of individuals in the respected class and

the mean value was recorded in Table 3.7. The bias values were calculated by finding the average difference of true proportion of individuals with estimated proportion of individuals. Comparable to parameter estimation in previous section, sample size and accuracy of classification is positively correlated, meaning as the sample size increases the accuracy of classification improves significantly. As the table indicates, the most accurate estimation was for the longest test with 35 items and three attributes. This indicates that the test length has a significant effect on classification as well as parameter recovery.

Table 3.7

Comparison of Classification for Various Generated Datasets

	N=10,000	N=5,000	N=1,000	N=500
		7 items with 3 attributes		
Correlation	0.96	0.96	0.82	0.69
Bias	0.0141	0.022	0.042	0.054
		35 items with 3 attributes		
Correlation	0.99	0.99	0.93	0.90
Bias	0.0040	0.0049	0.0095	0.016
		15 items with 4 attributes		
Correlation	0.98	0.96	0.80	0.68
Bias	0.0039	0.0054	0.013	0.019
		31 items with 5 attributes		
Correlation	0.98	0.96	0.82	0.73
Bias	0.0019	0.0027	0.006	0.008

Discussion

The main purpose of CDM’ss is to provide diagnostic information on examinees' mastery of a set or subsets of attributes. Diagnostic models aim at classifying examinees into one of number of pre-assigned mastery profiles. The purpose of this study was to contribute to a better understanding of the effects of sample size and number of attribute with regards to test length and item complexity on item parameter recovery and classification accuracy. To allow for all possible item complexities to be included in the dataset, the test lengths have been chosen accordingly.

This simulation study differed from previous ones in the type of model used, sample sizes, item complexity, test length with regards to attributes, and extended number replication. We have used the CRUM in a log-linear framework per of the trend in recent advances in cognitive models (e.g. Rupp, Templin, & Henson, 2010; von Davier, 2005). The sample sizes used included a wide range of 50 to 10,000 examinees to achieve broader generalizability of results. The test length was selected in a manner in which all possible combinations of attributes were included for variety of item complexity. The results have shown that much can be learned from investigating this model under various sample size, test length, attribute level, and item complexity.

The advantage of using the CRUM as a log-linear model is that as the number of items increase the number of item parameter increase, despite the similarity of the Q-Matrix specification. Hence, longer test lengths have more reliable parameter and classification recovery.

One of the general findings of this study is the effect of test length on the parameter and the classification accuracy: this was shown in datasets of seven versus 35 items with three attributes. Another general finding of this study is the effect of sample size on the item parameter and classification recovery. The reliability of the results increases with an increase in the sample size. Studying the effect of sample size on parameter estimation can help researchers learn about the dependability and reliability of item parameter estimation and examinees classifications. The findings of this study extend the existing study on parameter recovery such as Choi, Templin, and Cohen (2010) and Kunina-Habenicht (2010). These researchers' simulation study on a limited number of observations indicated that models such as the CRUM can accurately recover item parameters well with large datasets. While the previous studies have focused on a few numbers of datasets, this study has extended the number of items with attribute combinations, sample size, and item complexity for a greater generalizability of the findings.

The followings are a summary of the findings of this study with four rules of thumb suggestions for practitioners:

1. To have an accurate item parameter recovery and classification for datasets with five attributes and 31 items, with all possible item complexities included in the dataset, a minimum sample size of 500 or more is needed.
2. To have an accurate item parameter recovery and classification for datasets with four attributes and 15 items, with all possible item complexities included in the dataset, a minimum sample size of 5,000 or more is needed. The sample size can be decreased to 500 if the item complexity stays at two or less per item.
3. To have an accurate item parameter recovery and classification for datasets with three attributes and 35 items, with all possible item complexities included in the dataset, a minimum sample size of 50 or more is needed.
4. To have an accurate item parameter recovery and classification for datasets with three attributes and seven items, with all possible item complexities included in the dataset, a minimum sample size of 10,000 or more is needed. The sample size can be decreased to 5,000 if the item complexity stays at two or less per item.

This study's results and conclusions are limited to application of the CRUM in log-linear format and Mplus limitation in the estimation approach. Another limitation of this study is the number of attributes used, three, four, and five attributes was the extent of this study, which might not be practical in large standard test settings. This study utilized CRUM, which can be compared with other existing models. A comparison of EM algorithm with MCMC algorithm could help with consistency of the results as well as preference of one over the other with respect with accuracy and convergence time.

In future studies comparison of EM estimation versus MCMC algorithm would be of interest. The convergence of CDMs based on EM algorithm is another important difficulty facing those who pursue these methodologies. Investigation of comparison of MCMC algorithm versus EM can provide information on the appropriate usage of these algorithms and advantage and disadvantage of one versus the other. Future study can, also, be focused on extending these findings to various numbers of examinees and items as well as models with various test length with only one or two attributes per item instead of inclusion of all possible item complexities.

References

Choi, H., Templin, J., & Cohen, A. (2010). *The impact of model misspecification on estimation accuracy in diagnostic classification models.* Unpublished manuscript.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies on cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595-624.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 37*, 359–374.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Harre, R. (1970). *The principles of scientific thinking.* Chicago, IL: University of Chicago.

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

Harwell, M., Stone, C.A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

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

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Lighton, & M. J. Gierl (Eds.), *Cognitive assessment for education: Theory and applications* (pp. 19-50). New York, NY: Cambridge University.

Kunina-Habenicht, O. (2010). *Theoretical and practical considerations for implementing diagnostic classification models: Insights from simulation-based and applied research*. (Doctoral dissertation, Humboldt University of Berlin). Retrieved from http://edoc.hu-berlin.de/dissertationen

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81.

Lai, H., Gierl, M. J., & Cui, Y. (2012, May). Item consistency index: An index for cognitive diagnostic assessment. *Proceeding of* the *Centre for Research in Applied Measurement and Evaluation*, Vancouver, Canada. Retrieved from http://www2.education.ualberta.ca.ezproxy.lib.vt.edu:8080/educ/psych/crame/docs/April%202012/NCME%202012%20ICI.pdf

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*(4), 523-547.

Mislevy, R. J. (1995, May). *Probability-based inference in cognitive diagnosis*. Paper presented at the Office of Naval Research Contractors Conference, Iowa City, IA.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to Evidence-Centered Design*. (Technical Report No. 632). Retrieved from University of California, Center for the Study of Evaluation website: http://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Muthén, L.K., & Muthén, B.O. (2010). Mplus user’s guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Norris, S. P., Macnab, J. S., & Phillips, L. M. (2007). Cognitive modeling of performance on diagnostic achievement tests: A philosophical analysis and justification. In J. P. Lighton,

& M. J. Gierl (Eds.), *Cognitive assessment for education: Theory and applications*. (275-318). New York, NY: Cambridge University

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit sore scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53* (3), 315-333.

Patz, J. R. & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

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

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

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

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173-180.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354.

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

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

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

Chapter 4

Manuscript 2: The Accuracy of Relative Fit Indices in Detecting Misfits for Cognitive Diagnostic Model

Abstract

Given their novelty, little research has been directed toward cognitive diagnostic models (CDMs), specifically the Compensatory Reparameterized Unified Model (CRUM). The CRUM is both simple and accurate, but as of yet relative fit indices have not been evaluated for the model. Four CRUM simulation studies analyzed variable sample size and test length/ number of attribute combinations, for sensitivity in relative fit indices under two Q-Matrix misspecifications: (1) an overfitted Q-Matrix (2) a complete reverse Q-Matrix. The data were generated with the true Q-Matrix model then were analyzed with mis-specified Q-Matrices for comparison of fit indices performance. The three relative fit indices performance tested were AIC, BIC, ssaBIC. A direct relationship between research designs and performance of these relative fit indices was found. The global fit indices performed better for longer test length such as datasets with 35 items and 3 attributes. Initial findings hint at CRUM's promising future in the field of educational testing. The following study can be used for those in the field as a guide for CRUM in praxis.

Key words: CRUM, relative fit indices, Q-Matrix misspecification, cognitive diagnostic model

Cognitive diagnostic models (CDM) are probabilistic psychometric models designed to estimate a student's understanding of the measured items at skill/attribute level, as master or non-master of sets of attributes (de la Torre & Douglas, 2008). Their main purpose is to enable researchers to evaluate students' cognitive processes from their item response patterns. CDM allows for investigation of the existing relationship between an examinee's item responses and their underlying ability in a specific attribute. These models have been referred to by various terms such as cognitive diagnostic model, diagnostic classification model, log-linear model, latent response model, and restricted latent class model (Rupp, Templin, & Henson, 2010).

Conventional test theories such as item response theory (IRT) and classical test theory (CTT), used in most educational testing programs, are useful in picking out students most likely to succeed, but not for helping them to succeed (Mislevy, Almond, & Lukas, 2004; Norris, Macnab, & Phillips, 2007; Snow & Lohman, 1989). Cognitive diagnostic models, on the other hand, can help generate a detailed outline of the students' mastery or non-mastery of sets of attributes for the purpose of teaching and training (Huff & Goodman, 2007).

During the last two decades, there have been many approaches to cognitive diagnostic models. Von Davier (2005) and then later Rupp, Templin, and Henson (2010) have suggested unifying the existing models using a log-linear approach. Rupp, Templin, and Henson (2010) have an extensive review of log-linear models; Compensatory Reparameterized Unified Model (CRUM) utilized here is one of the models they have introduced. Log-linear diagnostic models share two common features: attributes and a Q-Matrix .

Attributes, or skills, are latent variables or constructs: they explain people's understanding of a subject matter. Attributes are not observed directly, rather inferred by a set of observations such as item responses (Borsboom & Mellenbergh, 2007). They link students' performance to their cognitive processes through the use of a Q-Matrix.

The Q-Matrix is the building block of CDMs and its accuracy is fundamental for validity of the results; its misspecification would create a misrepresentation of the examinees (Tatsuoka, 1983). Development of a correct Q-Matrix in praxis is not an easy task and many unpredictable mistakes can occur in the process. Prior knowledge of some consequences of these misspecifications can help researchers and practitioners plan in advance for extreme and not so extreme cases. An incorrect specification of Q-Matrix should reveal itself as model misfit. Cognitive diagnostic models in log-linear format are new psychometrics methodologies; despite increasing interest very little is known about their fit estimation and their accuracy under various research designs (Choi, Templin, & Cohen, 2010; Kunina-Habenicht, 2010; Rupp, Templin, & Henson, 2010; von Davier, 2005). CDMs in log-linear format unify the various models that exist; a single model instead of a multitude of independent models that then can use a common estimation algorithm.

Q-Matrix Misspecifications

Rupp and Templin (2008) suggested categorizing Q-Matrix misspecifications into two types: over and under specification. An over-specified Q-Matrix is defined as a Q-Matrix where an element of $\alpha_k=0$ has been replaced by $\alpha_k=1$, implying a more difficult item, hence more informative item, and an item parameter that should not exist is estimated. An under-specification of the Q-Matrix is defined as replacement of $\alpha_k=1$ with $\alpha_k=0$, implying an easier item, hence less informative item, and an item parameter that should exist is not estimated; indicating a model with bad fit.

Research on Relative Fit Indices

To date, there have only been a few studies concentrating on the effect of Q-Matrix misspecification on model fit: more specifically, there have only been two studies examining the sensitivity of the relative fit indices in log-linear models.

The most recent study was by Kunina-Habenicht, Rupp, and Wilhelm (2012) on examining the fit estimation of a log-linear model—the log-linear cognitive diagnostic model (LCDM)—with three versus five attributes and 1,000 versus 10,000 examinees with 150 replications. Their Q-Matrix misspecification included an over and under-specified Q-Matrix with item complexity limited to three attributes per items. Their results showed that the parameter estimation had a direct relationship with test length and sample sizes. For the larger sample sizes of 10,000 examinees, the main effect parameters and classification were estimated accurately and the model fit was correctly identified for all four specifications. They further argued that the item parameters were estimated more accurately for the intercept and main effect rather than the two and three-way interactions.

The second study was performed by Choi, Templin, and Cohen (2010) who analyzed the sensitivity of relative fit indices for deterministic inputs noisy and gate model (DINA), the CRUM, deterministic input noisy or gate model (DINO), and the log-linear cognitive diagnosis model (LCDM). They examined the accuracy of AIC and BIC for sample sizes of 100, 200, 500, 1,000, 2,000 and 4,000, four attributes, and 40 items, with 100 replications. Their simulated tests were limited to items with one or two attributes per items, hence, keeping the item difficulty and complexity simple.

Their results showed that AIC and BIC were sensitive to model misfit for samples of 200 or more examinees. The comparison suggested that the CRUM was one of the most robust models in parameter recovery, classification, as well as fit estimation, despite generating the data under log-linear cognitive diagnostic model (LCDM) specification. The global fit indices performed the best under CRUM specification and in terms of accuracy of parameter recovery, CRUM performed as well as LCDM under which the data was generated.

This study builds on the existing studies and aims at extending their findings to a greater generalizability using the model that has been suggested to perform well among other log-linear models. The CRUM has shown to be promising in accuracy and simplicity among other compensatory log-linear models. To do so, relative fit indices' sensitivity on detecting fit under Q-Matrix misspecification was analyzed using simulation methods. Three relative fit indices were studied: Bayesian Information Criterion (BIC) (Akaike, 1978), Akaike Information Criterion (AIC) (Akaike, 1973), and sample size adjusted BIC (ssaBIC) (Sclove, 1987).

The current study was designed to investigate the effects of severe Q-Matrix misspecification on relative model fit. Specifically, this study aimed to answer the following research questions:

1. How accurately do the three relative fit indices (AIC, BIC, sample size adjusted BIC) detect the fit of the model under various *test length/Q-Matrix* and *sample size conditions* when the dichotomous data is generated from a CRUM?
2. How do different relative fit indices such as AIC, BIC, and sample-adjusted BIC compare in detecting misfit under various conditions of model misspecification when the dichotomous data is generated from a CRUM?

Basic Structure of CDMs

One essential input for cognitive models is the building of the Q-Matrix, which was introduced by Tatsuoka (1983) as the loading matrix for test items with attributes as its element,

q_{jk}. The development of Q-Matrix requires expert knowledge on the topic, it relies on specialist judgment and its accuracy is fundamental on the results of CDM estimation (Tatsuoka, Birenbaum, & Arnold, 1989; Tatsuoka, 2002). The Q-Matrix specifies whether mastery of attribute *k* is necessary for an accurate response to item *i*, if so $q_{ik} = 1$ otherwise $q_{ik} = 0$.

For example, an item ($i = 1$) in a test with three attributes could require the first and second attributes for an accurate response, but has no relationship to attribute three. This is indicated by the following Q-Matrix entry: $q_{11} = 1, q_{12} = 1, q_{13} = 0$ or (1, 1, 0).These arrays of items by attributes constructs the Q-Matrix with *i* items and *k* attributes, below shows the Q-Matrix for the example discussed:

$$QMatrix = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

One common output for all CDMs is the classification of the individuals, posterior class membership, or classification of each examinee based on the examinee's item response pattern. Attributes and latent classes have an exponential relationship, meaning a linear increase in the number of attributes causes an exponential increase in the number of latent classes. For example a test with four attributes would have $2^4 = 32$ latent classes; an examinee's mastery profile could fall into one of these 32 latent classes. A test with five attributes would have $2^5 = 64$

Cognitive diagnostic models, regardless of their statistical approach, are categorized into two distinct diagnostic models of compensatory versus non-compensatory models. The distinction is in direct relationship with the attributes of the subject matter, the examinee's cognitive process, and the development of the Q-Matrix. If the attributes interact within the Q-Matrix in such a way that an accurate response to an item requires mastery of all involved attributes, the item falls into the category of non-compensatory: meaning the mastery of all attributes involved in that item is required for a successful response to that item. In other words the knowledge of one attribute cannot compensate for the lack of knowledge if the other.

In compensatory items, a successful application of each required attribute increases the probability of correct response to that item. In other words, the attribute specifications of each item interact in such a way that proficiency of one attribute can compensate for the non-proficiency of the other skills.

Log-Linear Re-parameterized Unified Model (CRUM)

Recently, cognitive diagnostic models (CDM) have been formulated into a more universal format for the purpose of generalizing the models (Roussos & Templin, 2007). The general class of general diagnostic models (GDM) was developed by von Davier (2005) to make the existing cognitive models similar to an IRT format. Rupp, Templin, and Henson (2010) developed a variation of the GDM under the title of log-linear cognitive diagnostic model (LCDM).

Existing studies have suggested that models with intercept and main effect parameters, similar to the CRUM, can recover their parameters more accurately than models with interaction parameters (e.g., Choi, Templin, & Cohen, 2010; Kunina-Habenicht, 2010). Despite its promising indications, there have not been any existing studies investigating CRUM’s validity, fit estimation, and applicability.

Among the existing compensatory models, the log-linear CRUM is one of the more flexible models; letting items with the similar Q-Matrix entries have different parameters (Henson, Templin, Willse, 2009; Rupp, Templin, Henson, 2010). The full log-linear representation of a CDM involves two components of probability and measurement. To formulate the probability of a item response, let items be represented by $i=1,....I$, respondents be represented by *r=1,…,R*, the person ability be represented by—latent classes—*c=1,….C*, and latent attributes be represented by *k=1,…..K*. Given these specifications, the CRUM will have the following formulation for item probability (Rupp, Templin, Henson, 2010):

$$\mathrm{\pi_{ic}} = \mathrm{P(X_{ic} = 1|\alpha_{ic})} = \frac{\exp(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)}\alpha_{c\alpha}q_{i\alpha})}{1 + \exp(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)}\alpha_{c\alpha}q_{i\alpha})}$$

In the formula pictured above, $\mathrm{\pi_{ic}}$, $P(X_{ic} = 1|\alpha_{ic})$, represents the probability of a correct response given the latent class $\mathrm{\alpha_{ic}}$ specification; π_{ic} represents the probability of an accurate response for an examinee in latent class *c;* $\lambda_{i,0}$ represents the intercept for item *i*, also referred to as *item difficulty*; $\lambda_{i,1,(a)}$ represents the slope or vector of attribute *a* from the Q-Matrix; X_{ic} represents the examinee's observed response to item *i*. This model allows for the intercept to change across the items and the slope to vary across items at the attribute level.

To formulate the measurement component, let μ_c represent the mixing proportion using marginal probability of the latent classes to calculate the measurement sub-formula υ_c:

$$\mu_c = \sum_{a=1}^{A} \gamma_{1,(a)}\alpha_{ca}$$

$$\upsilon_\mathrm{c} = \frac{\exp(\mu_\mathrm{c})}{\sum_{c=1}^{c} \exp(\mu_\mathrm{c})}$$

In the above formula, $\gamma_{1,(a)}$ represents the main effect's coefficient of regression for latent class a. Combining the structural sub-formula with measurement sub-formula will result in the following CRUM representation:

$$\mathrm{P(X_r = x_r)} = \sum_{c=1}^{C} \vartheta_\mathrm{c} \prod_{i=1}^{I} \pi_{ic}^{x_{ir}}(1 - \pi_{ic})^{1-x_{ir}}$$

In the above formula, $\pi_{ic}^{x_{ir}}$ represents the probability that examinee *r* correctly responds to item *i* and x_{ir} represents the observed response for examinee *r* to item *i*.

Method

It is important to note, that the models under study varied in attribute /item combinations with various complex item specifications, hence, differed in their designs and therefore, difficult to directly compare to each other. However, in a broader sense, test designs vary in practical situations and it is important to evaluate them under as many designs as possible. Furthermore, the focus is on providing information regarding the most extreme cases to be able to set a baseline for future designs. Since there is no existing study examining the behavior of CRUM under any degree of misspecification, it was important to evaluate two extreme cases as a baseline. This extreme misspecification increases the probability of fit detection and allows for a comparison among the three indices and understanding of their performance under various research designs. Given the attribute selections in this study, it would be hard to state that one design is consistently better than the other. Instead, it is argued that once we are familiar with the criteria of one set of attributes combinations, comparisons within each model can provide useful guidance for design selection.

The justification of such an approach is based on existing methodologies with conservative over-fitted or under-fitted QMatrices. Both Choi, Templin, and Cohen (2010) and Kunina-Habenicht, Rupp, and Willse (2012) studies have examined relatively simple Q-Matrices in which items measured two or three attributes, a percentage of the tests were changed to overfitted or underfitted Q-Matrix, and tests with long test lengths were examined. This study, on the other hand, aims at examining tests with short as well as long test length, all combinations of attributes in the Q-Matrix, and examining extreme Q-Matrices of complete reverse specification as well as over-fit specification. The goal was to examine the sensitivity of relative fit indices for the CRUM under theses extreme cases. There is an almost impossibly large number of ways a Q-matrix could be misspecified, from very minor (change in a small number of items) to major (change in most or all items). Almost none of this has been investigated, so a reasonable place to start is with major misspecification, meaning that if indices do not detect this, they won't detect any sort of misspecification.

This simulation study consisted of three sets of data generated from the CRUM model with consideration of incorporating all possible attribute combinations and one additional test, 35 items with 3 attributes, for examining the effect of test length. The first data was created with the true Q-Matrix specification, the second test was analyzed with complete reverse Q-Matrix specification, and the last test was analyzed with an over-fitted Q-Matrix specification. Table 4.1 shows the tests with true Q-Matrix specification for all tests used in this study. One note to mention is that tests with 3 attributes with 7 items, 4 attributes with 15 items, and 5 attributes with 31 items have every combination of the Q-Matrix entries with no repeat or missing entries.

Number of Examinees

One of the independent variables in this study was the number of examinees. Item response datasets were generated with sample sizes of 50, 100, 500, 1,000, 5,000, and 10,000 observations. Previous studies discussed above have shown that sample sizes can have a significant effect on parameter recovery and classification accuracy as well as fit estimation using relative fit indices (Choi, Templin, & Cohen, 2010; Kunina-Habenicht, Rupp, & Wilhelm's, 2012). This study focuses on a larger range of sample sizes, from 50 to 10,000 and attempts to establish a guideline for the purpose of using cognitive models, not only in a large scale setting, but in a regular classroom setting with the number of examinees as low as 50.

Number of Items with conjunction with attributes for Data Manipulations

Table 4.1

Simulated Tests with True Q-Matrix Specifications

J=7, K=3		J=15, K=4		J=31, K=5		J=35, K=3	
Items	Q-Matrix	Items	Q-Matrix	Items	Q-Matrix	Items	Q-Matrix
1	0 0 1	1	0 0 0 1	1	0 0 0 0 1	1	0 0 1
2	0 1 0	2	0 0 1 0	2	0 0 0 1 0	2	0 1 0
3	0 1 1	3	0 0 1 1	3	0 0 0 1 1	3	0 1 1
4	1 0 0	4	0 1 0 0	4	0 0 1 0 0	4	1 0 0
5	1 0 1	5	0 1 0 1	5	0 0 1 0 1	5	1 0 1
6	1 1 0	6	0 1 1 0	6	0 0 1 1 0	6	1 1 0
7	1 1 1	7	0 1 1 1	7	0 0 1 1 1	7	1 1 1
		8	1 0 0 0	8	0 1 0 0 0	8	0 0 1
		9	1 0 0 1	9	0 1 0 0 1	9	0 1 0
		10	1 0 1 0	10	0 1 0 1 0	10	0 1 1
		11	1 0 1 1	11	0 1 0 1 1	11	1 0 0
		12	1 1 0 0	12	0 1 1 0 0	12	1 0 1
		13	1 1 0 1	13	0 1 1 0 1	13	1 1 0
		14	1 1 1 0	14	0 1 1 1 0	14	1 1 1
		15	1 1 1 1	15	0 1 1 1 1	15	0 0 1
				16	1 0 0 0 0	16	0 1 0
				17	1 0 0 0 1	17	0 1 1
				18	1 0 0 1 0	18	1 0 0
				19	1 0 0 1 1	19	1 0 1
				20	1 0 1 0 0	20	1 1 0
				21	1 0 1 0 1	21	1 1 1
				22	1 0 1 1 0	22	0 0 1
				23	1 0 1 1 1	23	0 1 0
				24	1 1 0 0 0	24	0 1 1
				25	1 1 0 0 1	25	1 0 0
				26	1 1 0 1 0	26	1 0 1
				27	1 1 0 1 1	27	1 1 0
				28	1 1 1 0 0	28	1 1 1
				29	1 1 1 0 1	29	0 0 1
				30	1 1 1 1 0	30	0 1 0
				31	1 1 1 1 1	31	0 1 1
						32	1 0 0
						33	1 0 1
						34	1 1 0
						35	1 1 1

The second manipulated factor in this study is the number of items in conjunction with the number of attributes. Research in CDM has consistently indicated that the number of items has significant effect on the parameter estimation and classification accuracy, as well as fit estimation accuracy (Henson, Templin, & Willse, 2009; Tatsuoka, 1990; Templin, Henson, Rupp, Jang, & Ahmed, 2009).

This researcher has investigated a wide range of item/attribute combination to evaluate the most extreme cases that could exist in a tests with three attributes with seven and 35 items, four attributes and 15 items, and five attributes with 31 items (Table, 4.1).

Rupp and Templin (2008) stated that three, four, and five attributes are considered as a norm for log-linear models. They further argued that the number of latent classes increases exponentially with the number of attributes, and these estimations are computationally very demanding. For these reasons, most researchers have limited their study to less than seven attributes and mostly at less than five attributes per dataset (e.g. Hartz, 2002; Maris, 1999; Templin & Henson, 2006).

Existing studies have been limited to tests with as few as 15 items with four attributes by Templin, Henson, Rupp, Jang, and Ahmed (2009) with the DINA model and as large as 50 items with five attributes by Kunina-Habenicht, Rupp, and Willse (2012) with the LCDM approach. While there have been studies with varying numbers of test lengths, none has considered including the most complex items, four and five attributes per item, and none has examined a test length of less than 15 items (e.g., Templin, Henson, Rupp, Jang, & Ahmed, 2009). Research studies have not yet examined comprehensive sets of datasets to examine the CRUM model.

Similarly, other studies included a limited and conservative item specification such as: Henson, Templin, and Willse's (2009) study of datasets with 40 items and seven attributes, Templin, Henson, Rupp, Jang, and Ahmed's (2009) study of datasets with 20 items and four attributes, and von Davier (2005) study of dataset with 40 items and four attributes.

Item Difficulty Specification for Data Generation

To simulate the data two main distributions were required, the person's ability and item difficulty. A purposeful distribution for the item difficulty was selected. An overall/baseline item difficulty in log linear model is related to the intercept of the item (λ_0); indicating the probability of responding to an item correctly if the respondent does not have any of the required skills for that particular item (Rupp, Templin, & Henson, 2010). The more attributes needed for a correct response to an item the more difficult that item tends to be (Rupp, Templin,& Henson, 2010).

Table 4.2

Parameter Assignment Used in Simulation Design

Item	skill1	skill2	skill3	skill4	skill5	λ0	λ1_1	λ1_2	λ1_3	λ1_4	λ1_5
1	0	0	0	0	1	-1.50	0	0	0	0	2
2	0	0	0	1	0	-1.50	0	0	0	2	0
3	0	0	0	1	1	-2.25	0	0	0	2	2
4	0	0	1	0	0	-1.50	0	0	2	0	0
5	0	0	1	0	1	-2.25	0	0	2	0	2
6	0	0	1	1	0	-2.25	0	0	2	2	0
7	0	0	1	1	1	-3.38	0	0	2	2	2
8	0	1	0	0	0	-1.50	0	2	0	0	0
9	0	1	0	0	1	-2.25	0	2	0	0	2
10	0	1	0	1	0	-2.25	0	2	0	2	0
11	0	1	0	1	1	-3.38	0	2	0	2	2
12	0	1	1	0	0	-2.25	0	2	2	0	0
13	0	1	1	0	1	-3.38	0	2	2	0	2
14	0	1	1	1	0	-3.38	0	2	2	2	0
15	0	1	1	1	1	-5.06	0	2	2	2	2
16	1	0	0	0	0	-1.50	2	0	0	0	0
17	1	0	0	0	1	-2.25	2	0	0	0	2
18	1	0	0	1	0	-2.25	2	0	0	2	0
19	1	0	0	1	1	-3.38	2	0	0	2	2
20	1	0	1	0	0	-2.25	2	0	2	0	0
21	1	0	1	0	1	-3.38	2	0	2	0	2
22	1	0	1	1	0	-3.38	2	0	2	2	0
23	1	0	1	1	1	-5.06	2	0	2	2	2
24	1	1	0	0	0	-2.25	2	2	0	0	0
25	1	1	0	0	1	-3.38	2	2	0	0	2
26	1	1	0	1	0	-3.38	2	2	0	2	0
27	1	1	0	1	1	-5.06	2	2	0	2	2
28	1	1	1	0	0	-3.38	2	2	2	0	0
29	1	1	1	0	1	-5.06	2	2	2	0	2
30	1	1	1	1	0	-5.06	2	2	2	2	0
31	1	1	1	1	1	-7.59	2	2	2	2	2

The purposeful assignment allows for a calculated difference between complex items with many attributes involved and simple items with small number of attributes involved. Hence, items with more involved attributes tend to have larger intercepts, indicating difficult items, and items with less involved attributes tend to have lower intercepts, indicating easier items. On the other hand, the main effect parameters (λ_1) were assigned an equal weight, indicating equality of difficulty for all involved attributes, these values are similar to Rupp, Templin, and Henson's (2010) suggestions. Table 4.2 shows items one, two, four, and eight, have the lowest absolute value of -1.5 for their intercept—easy items—while item 31 has the highest absolute value of -7.59—difficult item.

Person Ability Specification for Data Generation

Another important factor in creating the simulated data is determining the person ability distribution, indicated by the mastery attribute profile, or latent class membership (Kunina-Habenicht, 2010; Rupp, Templin, & Henson, 2010). To eliminate any biased priori, locating examinees differently in the latent classes, uniform distribution was selected, though Mplus specification requires the user to assign the last latent class specification to be 0 (Table 4.3). It is customary to assume uniform distribution if the researcher does not have a clear understanding of the population distribution: this avoids biased estimation and the estimation will not require modification of the results to suit the population or estimation itself (Gewekei, 1989; Qian, Stow, & Borsuk, 2003). Since these models are new and have not been frequently tested in praxis, the distribution of the examinees was not truly uniform distributions rather were based on existing studies suggestions. The population of examinees in the first and the last latent class was the highest grounded in the previous studies on existing data (eg. Haertel, 1989; Von Davier, 2005) that suggested the first and last latent class has higher proportion of examinees than the remaining latent classes.

Table 4.3

Class Membership Proportion for Simulation Design

υ_c/Prob	Class	μ_c
0.2377	1	0.00
0.0874	2	1.00
0.0874	3	1.00
0.0874	4	1.00
0.0874	5	1.00
0.0874	6	1.00
0.0874	7	1.00
0.2377	8	0.00

Summary of Simulation Methodology

Two main programs of SAS version 9.2 and Mplus version 5 (Muthen & Muthen, 2010) were used for this study. Data was simulated for the CRUM model and the above-mentioned specifications using SAS. After generating the simulated data, the generated data was input into Mplus for parameter estimation and replicated 200 times for each cell in the design to eliminate any biased estimation.

Table 4.4

Simulated Experimental Units, the Design Summary

Manipulated Parameter /Treatment	Number of manipulation/ Treatment	Assigned value
Number of observations	6	N=50, 100, 500, 1,000, 5,000, & 10,000
Number of attributes	3	k= 3, 4, 5
Number of items-attributes	4	j= 7-3, 15-4, 31-5, 35-3
Number of replications		200
Number of simulation studies		6*4=24
Q-Matrix misspecification of complete reverse	24	6 observation levels, 4 item/attribute levels
Q-Matrix misspecification of over-fit	18	6 observation levels, 3 items/attribute levels
Q-Matrix True specification	24	6 observation levels, 4 item/attribute levels
Total models generated		66
Total number of generated data		200 * 66 = 13,200

Table 4.4 shows a summary of data simulation with regards to items, attributes, and number of examinees. A total of six observations and four test length/attribute combinations, under true Q-Matrix, complete opposite Q-Matrix, and over-fitted Q-Matrix—excluded five attributes and 31 items—were simulated. These generated data were replicated 200 times for accuracy estimation, hence 200 x 66 = 13,200 data were generated.

This study generated 200 replications for estimation accuracy; Choi, Templin, and Cohen's (2010) study generated 100 replication for estimation accuracy; and Kunina-Habenicht, Rupp, Wilhelm's (2012) study generated 150 replications for estimation accuracy. As Templin and Henson (2006) discovered, these models, specifically in Bayesian framework, can take several hours or even days to convergence.

Additionally, Harwell, Stone, Hsu, and Kirisci (1996) focused on the number of necessary replications for reliable and stable results from Markov chain Monte Carlo's (MCMC) studies and recommended a minimum of 25 replications. They further stated that the number of replications needed changes based on the complexity of the model; models with higher dimensions and numerous effects require more replications (e.g., main effects and interaction effect). The model studied here is based on main effect only, which can be estimated reliably with 150 or 200 replications.

Model evaluation Approach

To evaluate the relative fit indices the average bias was estimated, by calculating the average differences between the base model with correct Q-Matrix specification and the model with misspecified Q-Matrix:

$$avg\ Bias = \frac{1}{r}\sum_{r=1}^{200}(\hat{\tau}_r - \tau_r)$$

In the above formula, *r* represents the number of replications, $\widehat{\hat{\tau}_r}$ represents the estimated fit indices under misspecified Q-Matrix, and τ_r represents the estimated fit indices under true Q-Matrix specification. Any positive value would indicate accurate fit estimation of the relative fit indices. Another evaluation approach used in this study was the calculation of the percentage of inaccurate model estimation:

$$Missed\ Bias = \frac{missed}{r} * 100$$

In the above formula, *missed* represents the count for the number of times the model failed to identify fit, meaning the relative fit indices were higher for the model with true Q-Matrix, and *r* represents the number of replications.

The first research question focuses on the performance of the relative fit indices: AIC, BIC, and ssaBIC. To calculate these three relative fit indices the likelihood is estimated (G^2). The followings are the formula for AIC, BIC, and ssaBIC:

$$G^2 = -2ln$$

$$AIC = G^2 - 2df$$

$$BIC = \ G^2 - df * [\ln(N)]$$

$$ssaBIC = \ G^2 - df * [\ln\left(\frac{(\mathrm{n}\ +\ 2)}{24}\right)]$$

Results

Relative Fit Estimation

The first research question was on AIC, BIC, and ssaBIC’s accuracy of detecting misfit under various Q-Matrix misspecifications and research designs. Table 4.5 summarizes the performance of the relative fit indices under the various research conditions. At first glance, it is clear that the larger the sample size, the number of items with regards to the number of attributes, and the simpler the item complexity— simple items are items with few attributes—the more accurate the relative fit indices have performed in detecting misfit.

An inspection of the bias for the three relative fit indices on the test of five attributes with 31 items indicated that these relative fit indices could detect misfit with 100% accuracy for datasets of N ≥ 100, while the accuracy decreases to 50% when N=50 (Table 4.6). Similar to the previous dataset, the three relative fit indices were able to detect misfit with 100% accuracy for datasets of 15 items with four attributes for N ≥ 100; although the findings from previous study suggested a questionable item parameter recovery for datasets of N≤100. This could be based on the high degree of Q-Matrix misspecification (Galeshi & Skaggs).

Table 4.5

Comparison of relative fit: Reverse & Over-fit Q-Matrix misspecification

	N=10,000		N=5,000		N=1,000		N=500	
Test with 31 items and 5 skills item specification: i=5 K=1, i=9 K=2, i=10 K=3, i=5 K=4, i=2 K=5 (Over-fit)								
Bias AIC	18,272.30		9,115.00		868.6		666.6	
Bias BIC	18,248.40		9,082.40		842.5		666.1	
BiasssaBIC	18,258.20		9,098.30		858.4		662.5	
AIC %	0%		0%		0%		0%	
BIC %	0%		0%		0%		0%	
SsaBIC %	0%		0%		0%		0%	
Test with 15 items and 4 skills item specifications: i=4 K=1, i=6 K=2, i=4 K=3, i=1 K=4								
	Opposite	Over-fit	Opposite	Over-fit	Opposite	Over-fit	Opposite	Over-fit
Bias AIC	6,744.60	49.35	3,374.70	47.95	666	35.2	333.4	32.32
Bias BIC	6,715.80	231.24	3,348.60	220.43	646.4	180.6	316.7	153.3
BiasssaBIC	6,728.50	142.26	3,361.30	131.45	659.1	91.7	329.3	64.5
AIC %	0%	0%	0.67%	0%	0.67%	2%	0.67%	2%
BIC %	0%	0%	0%	0%	0%	0%	0.67%	0%
SsaBIC %	0%	0%	0%	0%	0%	0%	0.67%	0%
Test with 35 items and 3 skills item specification: i=15 K=1, i=15 K=2, i=5 K=3								
	Opposite	Over-fit	Opposite	Over-fit	Opposite	Over-fit	Opposite	Over-fit
Bias AIC	38,189.50	1542.7	19,092.80	1161.4	3,532.50	298.3	1,868.40	169.1
Bias BIC	38,110.30	1867.25	19,021.20	1454.7	3,527.10	519.1	1,822.10	358.7
BiasssaBIC	38,145.20	1724.24	19,056.10	1311.7	3,529.50	376.2	1,857.10	215.9
AIC%	0%	0%	0%	0%	0%	0%	0%	0%
BIC%	0%	0%	0%	0%	0%	0%	0%	0%
SsaBIC %	0%	0%	0%	0%	0%	0%	0%	0%
Test with 7 items and 3 skills item specification: i=3 K=1, i=3 K=2, i=1 K=3								
	Opposite	Over-fit	Opposite	Over-fit	Opposite	Over-fit	Opposite	Over-fit
Bias AIC	325	8.9	174.4	7	34.1	5.4	24.4	5.3
Bias BIC	311.6	73.8	160.6	65.7	31.3	49.6	21.3	43.2
Bias-saBIC	320.4	45.2	169.1	37.1	33	21	23.2	14.7
AIC%	0%	2.70%	0%	7%	0.70%	16%	19.30%	15%
BIC%	0%	0%	0%	0%	0.70%	0%	19.30%	0%
SsaBIC %	0%	0%	0%	0%	0.70%	0%	19.30%	0.70%

The analysis of test with three attributes and 35 items indicated that the fit estimation was accurate for test as low as 50 observations—depending on the relative fit indices used, which will be discussed shortly—suggesting an increase in the test length can improve the fit detection. This knowledge can be helpful for small classroom settings. An interesting finding is the analysis of the test with seven items, three attributes, and inclusion of complex item suggesting that BIC and ssaBIC can accurately detect fit if the sample size is larger than 1,000.

Table 4.6

Comparison of relative fit: Reverse & Over-fit Q-Matrix misspecification small sample sizes

	N=100		N=50	
I= 31 K = 5				
Bias AIC	147.0		-57.2	
Bias BIC	134.0		-85.4	
BiasssaBIC	149.8		-81.7	
AIC %	0		50	
BIC %	0		50	
SsaBIC %	0		50	
I = 15 K = 4				
	Opposite	overfit	Opposite	Overfit
Bias AIC	231.2	17.4	37.9	21.2
Bias BIC	29.4	90.4	25.2	74.8
BiasssaBIC	142.3	1.9	32.8	-13.5
AIC %	0	11.33333	0%	7.0
BIC %	0	0	4%	0.0
SsaBIC %	0	47.3	0%	85.0
I = 35 K = 3				
	Opposite	Overfit	Opposite	Overfit
Bias AIC	367.5	253.5	176.9	57.8
Bias BIC	332.9	368.8	142.3	143.8
BiasssaBIC	361.6	232.8	173.4	2.5
AIC %	0%	0%	0%	1%
BIC %	0%	0%	0%	0%
SsaBIC %	0%	1%	0%	35%

Table 4.5 shows as a general rule that BIC is more accurate in detecting misfit, but the consistency varies slightly within different specifications. On the other hand (Table 4.6), the sensitivity of the relative fit indices in detecting fit for small sample sizes and complex item designs is consistent with the findings of Choi, Templin, and Cohen (2010) who suggested that the CRUM was the most accurate model in estimating fit, for sample sizes as small as 200. On the other hand, Kunina-Habenicht, Rupp, and Whilhelm’s (2012) study of the LCDM with two

and three-way interaction parameters suggested that observations of 1,000 had estimated fit with less than 95% accuracy.

Determining the Most Accurate Fit Indices

The second research question focused on the comparison of the AIC, BIC, and ssaBIC relative fit indices. Table 4.6 shows that AIC tends to identify the misfit in some cases, similar to the previous studies by Choi, Templin, and Cohen (2010) and Kunina-Habenicht, Rupp, and Whilhelm (2012). However, BIC tends to detect misfit more accurately, specifically for overfit Q-Matrix misspecification and shorter test lengths. These findings are different from the previous findings and suggest that a difference in the model used for data generation and analysis could have an effect on relative fit indices' sensitivity.

Discussion

This study's focus was on one of the more promising models, the CRUM, and its findings extend from simple to complex Q-Matrix specifications. The aim was to discover the CRUM's behavior under simple versus extreme research designs. Previous research on log-linear models had suggested that the CRUM can be more promising in detecting fit under Q-Matrix misspecifications.

A limitation of this study is the usage of simulation methodology: it is hard to validate the results without implementing them in practice. Another empirical limitation is that these models are hard to apply in practice. Tests with extensive content domains encompass many attributes: many attributes requires exponential number of latent classes to estimate and long test length that are impossible to administer.

The number of items, attributes, and examinees are limited and does not consider all practical designs. Another limitation is the Q-Matrix misspecification design: this study considered only two of many possible Q-Matrix misspecifications. The degree of misspecification is another limitation of this study. The number of items in the test can vary from one item to all items misspecifications. The only model that was considered in this study was the CRUM, which does not include all possible cognitive diagnostic models.

The results of this study suggests that AIC can detect fit more accurately than the other two fit indices when the degree of misspecification is opposite Q-Matrix and BIC can detect fit more accurately when the degree of misspecification is an overfit Q-Matrix. This finding contrasts with the two previous studies (Choi, Templin, & Cohen, 2010; Kunina-Habenicht, Rupp, & Wilhelm, 2012) which claimed that the accuracy of AIC surpassed the other fit indices in all cases they studied. The results indicated that ssaBIC was the least accurate among the relative fit indices.

Based on these findings, the best suggestion is to utilize the two criteria of AIC and BIC together. This approach provides reassurance on the robustness of the choice. This findings are similar to Kuha's (2004) who had suggested to use both AIC and BIC for reassurance of the fit, meaning unless one is suggested for a specific model make sure both indices suggest fit.

Another important finding of this investigation was the accuracy of global fit indices in detecting for a short test with seven items and three attribute. The results suggested that it is possible to detect misfit for such a short dataset if the number of examinees is larger than 5,000. One of the findings of this study that can be valuable for school administrators and teachers who would wish to assess student's attribute proficiency in an ongoing basis is the results of the tests of 15 items with four attributes. Educators can use the results from such tests to focus on a few skills at a time: they can assess student's progress for future instructional direction. The results suggest that test with 15 items and four attributes can be examined for misfit if the number of examinees is 500 or more. Same conclusion can be made for tests with five attributes and 31 items, teachers and schools can assess grade levels of their progress and design a more suitable and attainable curriculum based on that understanding. Testing student's progress on sets of attributes before moving to the next sets of attributes can help students build the ground necessary for understanding of the future subject. To summarize the findings of this study the following rules of thumb can be considered when performing a CDM analysis with sever misfit:

1. CRUM can be considered a robust model for fit estimation.
2. To reassure the accuracy of fit, it is recommended to use both AIC and BIC simultaneously to ensure the accuracy of the fit estimation.
3. It is possible to detect severe misfit for short assessment with three attributes and seven items, if the number of examinees is 5,000 or higher.
4. It is possible to detect severe misfit for test length of 15 items with four attributes, if the number of examinees is 500 or higher.
5. For samples of 50, a test of three attribute and 35 items are the minimum requirement.

More studies are required to support these findings. The CRUM needs to be tested in a practical setting to evaluate its performance under various research designs. Although, we suggest the use of the CRUM in the above specifications, some subject matters do require the use of non-compensatory models: it is important to evaluate the CRUM with other non-compensatory models for more practical use of CDMs.

Some other suggestions for future research is examining misfit with various degree of misspecifications with fixed or random misspecifications. Comparing the results with sever specifications can help practitioners with guidelines regarding the effect of Q-Matrix misspecifications.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics, 30*, 9-14.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. Lighton, & M. J. Gierl (Eds.), *Cognitive assessment for education: Theory and applications* (pp. 85-115). New York, NY: Cambridge University.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Choi, H., Templin, J., & Cohen, A. (2010). *The impact of model misspecification on estimation accuracy in diagnostic classification models.* Unpublished manuscript.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies on cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595-624.

Galeshi, R., & Skaggs, G. (2012). *Item parameter estimation accuracy with a cognitive diagnostic model, CRUM*. Manuscript in preparation.

Gewekei, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica, 57*(6), 1317-1339.

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

Harwell, M., Stone, C.A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Henson, R., Templin, J., & Douglas, J. (2007). Use of subscores for estimation of skill masteries. *Journal of Educational Measurement, 44,* 361-376.

Huff, K. & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Lighton, & M. J. Gierl (Eds.), *Cognitive Assessment for Education: Theory and Applications*. (pp. 19-50). New York, NY: Cambridge University.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods Research, 33* (2), 188-229. doi: 10.1177/0049124103262065

Kunina-Habenicht, O. (2010). *Theoretical and practical considerations for implementing diagnostic classification models: Insights from simulation-based and applied research*. (Doctoral dissertation, Humboldt University of Berlin). Retrieved from http://edoc.hu-berlin.de/dissertationen

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187-212.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to Evidence-Centered Design*. (Technical Report No. 632). Retrieved from University of California, Center for the Study of Evaluation website: http://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Muthén, L.K., & Muthén, B.O. (2010). Mplus user’s guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Norris, S. P., Macnab, J. S., & Phillips, L. M. (2007). Cognitive modeling of performance on diagnostic achievement tests: A philosophical analysis and justification. In J. P. Lighton, & M. J. Gierl (Eds.), *Cognitive assessment for education: Theory and applications*. (275-318). New York, NY: Cambridge University

Qian, S. S., Stow, C. A., & Borsuk, M. E. (2003). On the Monte Carlo methods for Bayesian inference. *Ecological Modelling, 15,* 269-277.

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

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

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

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Eds.), *Educational measurement* (pp. 263–331). New York, NY: Macmillan.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Applied Statistics*, *51,* 337-350.

Tatsuoka, K. K., Birenbaum, M., & Arnold, J. (1989). On the stability of students' rules of operation for solving arithmetic problems. *Journal of Educational Measurement, 26*(4), 351-361.

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

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

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

Chapter 5

Discussion

The main objective of this dissertation was to investigate two aspects of the new psychometric methodology for using cognitive diagnostic models (CDM). The dissertation focused on a more flexible and promising variation of CDM known as compensatory reparameterized unified model (CRUM). In addition to a summary of the research and findings, this chapter focuses on the limitations of this CRUM study and future direction for CDMs. It is organized by the following chapter subheadings: scope of the research, summary of the main results, results of the first study, results of the second study, rules of thumb, limitations of the study, limitations of the model, and future direction.

Scope of the Research

Rather than measuring general abilities—such as with IRT models—cognitive diagnostic models have been developed with the aim of identifying the presence or absence of multiple skills/attributes required for accurate response on test items. As a technique, CDM combines cognitive psychology with statistical test analysis to measure students understanding of sets of attributes used as formative assessments, CDM-based test could enable teachers to design instruction according to their student's learning process, remediate students' non-mastered attributes, and provide instructions to retain students' mastered attributes. Teachers can use cognitive feedback to boost students' awareness of their learning approaches and evaluate their effectiveness. Students could, with a detailed description of their attribute mastery, take control of their learning by targeting case specific attributes.

One motivation for evaluating tests with a CDM is that these models provide more detailed information than other existing psychometric models: they allow for analysis of items at the attribute level rather than the item level. In other words, each item is characterized by a particular set of attributes identified by its Q-Matrix vector of binary elements. Each item indicates presence or absence of attributes under investigation. Individual performance on the test item will thus indicate one's mastery or non-mastery of the measured attributes.

With the novelty of CRUM approach comes many unexplored issues that deserve special attention. So in order to investigate CRUM's application issues, two individual studies were derived and performed from the following theoretical research questions:

1. How accurately does the model recover its parameters under various test length/ Q-Matrix and sample size conditions when dichotomous data is generated from a CRUM?
2. How accurately does the model classify students into their true class of attribute mastery under various test length/ Q-Matrix and sample size conditions when dichotomous data is generated from a CRUM?
3. How accurately do the three relative fit indices (AIC, BIC, sample size adjusted BIC) detect the fit of the model under various *test length/Q-Matrix* and *sample size conditions* when the dichotomous data is generated from a CRUM?
4. How do different relative fit indices such as AIC, BIC, and sample-adjusted BIC compare in detecting misfit under various conditions of model misspecification when the dichotomous data is generated from a CRUM?

The first study was derived from questions (1) and (2) and tested the accuracy of CRUM's parameter recovery under true Q-Matrix specification. The second study was derived from questions (3) and (4) and tested the relative fit indices' sensitivity in CRUM to severe Q-Matrix misspecifications.

Summary of the Main Results

In the first study, the author generated 48,000 datasets with accurate Q-Matrices and examined the parameter recovery and the classification accuracy of CDMs in log-linear framework throughout various research designs.

In the second study, the author simulated 66,000 datasets run through accurate, under- and over-specification of the Q-Matrix. The author calibrated the accuracy of the relative fit indices and compared them with two various forms of CRUM Q-Matrix misspecifications within the log-linear framework.

Based on both the number of examinees and attributes, practitioners can use the findings from the two studies to decide on test items needed to accurately estimate a particular assessment. Because most existing research (e.g. Choi, Templin, & Chohen, 2010; Kunina-Habenicht, Rupp, & Wilhelm, 2012) has focused on large number of examinees, their findings are harder to implement in smaller classrooms, whereas the author's findings are not.

Results of the First Study

Implementing the CRUM for test evaluation necessitates an investigation of its performance throughout several, predetermined and designed simulations. In the first study there were twenty-four models replicated two hundred times (i.e. one "simulation"). Each simulation had sample sizes of 50, 100, 500, 1,000, 5,000, and 10,000. Each simulation also used a Q-Matrix with the following specifications: three attributes with seven and thirty-five items; four attributes and fifteen items; and five attributes with thirty-one items.

Similar to the existing studies, the results suggested a strong relationship between sample sizes and accuracy of item parameter recovery as well as classification accuracy. Parameter recovery indicated that both the intercept and the main effects were accurately recovered under large datasets of 10,000 regardless of test length and attributes per items, yet decreasing the number of items while simultaneously increasing the number of attributes compounded the difficulty of the parameter estimation accuracy. The same result held true when increasing the complexity of the items. That is, an increase in the number of attributes per item, making them more difficult, caused a less accurate parameter recovery with higher SE, specifically for shorter datasets. It needs to be mentioned that the results apply to the EM algorithm as implemented in MPlus.

Upon further examination, the analysis revealed a strong correlation between item complexity and classification accuracy even for datasets as short as seven items. For items with one or two attributes, intercepts and main effects were accurately and reliably estimated even in $N = 500$ conditions. For items with three, four, and five attributes, the estimation of parameters were unstable and unreliable. Although, the author was able to suggest a few guidelines, there was no clear “minimum” to the number of observations required for achieving an accurate and reliable parameter recovery. Summarily, the author’s findings align themselves with the findings of previous studies such as Kuninin-Habinich (2010), Rupp and Templin (2005), and von Davier (2005).

Results of the second study

Research questions (3) and (4) concentrated on the evaluation of the relative fit indices. Mainly, AIC, BIC, and ssaBIC indices were evaluated for detecting misfit. Their individual performances were then compared. The author took a reverse misspecification of the Q-Matrix and an over-fit specification of the Q-Matrix and compared them with the true specification of the Q-Matrix. By this measure, the indices were comparatively evaluated for their sensitivity (e.g. the low value of the relative fit indices indicated a better fitting model).

This simulation study extended two existing studies' evaluation of log-linear CDM fit indices. One study by Choi, Templin, and Cohen (2010) showed that relative fit indices can successfully detect misfit if the number of examinees are larger than 200 observations and the complexity of the items stays at maximum of two attributes per item. Similar to the previous study, Kunina-Habenicht, Rupp, and Wilhelm's (2012) study focused on simulated tests with LCDM specification. Their investigation of 1,000 and 10,000 examinees revealed that AIC and BIC could detect fit for both datasets given that the complexity of the datasets stays at three attributes per item. However, findings by both research groups indicate that sample size can affect the performance of relative fit indices in detecting misfit.

Historically, an empirical evaluation of the fit indices is crucial to the field, CDM, and CRUM since it is difficult to evaluate the accuracy of the Q-Matrix in praxis. Q-Matrix specification relies too heavily on the consensus expertise because it *assumes* categorical stability of the attributes and the corresponding test items. Given the importance of Q-Matrix specification on CDM estimation accuracy, evaluating the model's fit sensitivity under Q-Matrix misspecifications is vital: Q-Matrix reveals the harmonies and incongruences between theoretical hypothesis and the cognitive processes involved in item creation and response.

The major finding of this study was twofold. The performance of relative fit indices was found under various research designs. The sensitivity was also found by way of an evaluative comparison. AIC, BIC, and ssaBIC information indices were found to be sensitive to extreme Q-Matrix misspecifications. It revealed that datasets with large numbers of examinees, 5,000 or higher, can be evaluated for misfit even with as complex items as five attributes per item. Therefore, the study offers a promising, if not simplified, method for calculating and evaluating log-liner CRUM fit indices.

The comparison of the relative fit indices revealed that ssaBIC is the least accurate indices, while both AIC and BIC are the most accurate ones to implement. The unreliable fit detection in the simulation study illustrates that relative fit indices sensitivity can be substantively impaired for datasets with complex items, especially in applications with small sample sizes. Interestingly, fit estimation was accurate for datasets with small number of attributes and long test length despite item's complexities (e.g. datasets with three attributes and 35 items). This suggests that to compensate for small sample sizes a longer test can be administered for an accurate fit estimation.

Rules of Thumb

After examining a wide range of variables—sample sizes, attribute/test length, and item complexity—in log-linear CRUM estimation, the following generalizations can be offered:

- Datasets with five attributes and 31 items must have at least 500 observations for accurate parameter and classification estimation
- Datasets with four attributes and 15 items must have at least 500 observations for accurate parameter and classification estimation.
- Datasets with three attributes and 35 items must have at least 50 observations for accurate parameter estimation.
- Datasets with three attributes and seven items must have at least 5,000 observations for accurate parameter estimations.
- For the purposes of fit estimation, AIC and BIC relative fit indices are best; specifically used together for assurance.
- If test length is larger than 31 items for three and five attribute datasets, 100 examinees are enough for accurate fit estimation.

Limitations of the Study

The present study has at least six specific limitations. First, the CRUM does not include two or three way interaction parameters that could affect the estimation accuracy. Although one can infer from studies that the CRUM can be implemented with a high degree of accuracy, there remains only one such study, Choi, Templin, and Cohen (2010), comparing it to other models. Second, the study investigates dichotomous data, or, to rephrase, the data does not include partial credit and open ended response items. This, of course, limits the generalizability of the findings. Third, distractors were not included. As the reader may know distractors can add valuable insight to the student's attribute mastery profile and to the validity of the findings. Fourth, the study does not include empirical data but rather simulated data. Fifth, the study does not compare equal number of items, but rather number of items varies with the number of attributes, making the comparison of the result more difficult. Sixth, the study does not include all possible Q-Matrix misspecifications, but rather implements only two Q-Matrix severe misspecifications of over- and reverse-specifications of the Q-Matrix, affecting the generalizability of the findings.

Limitations of the Model

One major limitation of cognitive diagnostic models is the difficulty in implementing the methodology to exams that are intended to measure wide-range of content domains. Such tests would involve a wide range of attributes and hence would need a test with many questions, making them impossible to administer. Tests that measure a narrow content area where each attribute is measured by multiple items are the most appropriate for cognitive model analysis. High representation of the attributes is needed for accurate estimation: this increases the degree of freedom by providing adequate information for each attribute.

At the same time, increasing the number of attributes creates models for which item parameters are very difficult to estimates. Additionally, a linear increase in the number of attributes exponentially increases the number of latent classes. This increase in the number of latent classes creates sparse cells, making their estimation and fit analysis very difficult.

Another limitation for application of CDMs is the lack of availability of software. Although, there are a few freeware applications, such as von Davier (2005) *mdltm*, Henson's *LCDM* and one commercially available software, *Arpeggio* and *MPlus* syntax-based, their reliability and validity has not yet been tested extensively.

Future Direction

To extend this study's findings a comparison of CRUM with other cognitive models specifically a non-compensatory model such as DINA, NIDA, or N-CRUM is recommended. This comparison would create a deeper understanding of cognitive models with various designs for implementation in praxis. Also extending the number of items and skills would allow for greater generalizability. As informative as this study was, it could not possibly encompass all possible designs and a future extension of this study by inclusion of many more research design can add to the validity and reliability of these findings.

One future direction would be retrofitting an existing test to one of the six diagnostic models, evaluating their item parameter estimation, accuracy of examinees' classifications, and sensitivity of relative fit indices. Although implementing CDM empirically seems promising, retrofitting has its own criticism. Retrofitting existing tests to CDM structure can result in misdiagnosing student's attribute mastery (Gierl, cui, & Zhou, 2008). Another approach would be developing tests that are designed with CDM specifications. These CDM-designed tests can then be analyzed for the accuracy of item parameter recovery, examinees classification, and the sensitivity of relative fit indices in detecting misfit.

Despite its obvious effectiveness for assessment purposes, diagnostic models have not received the deserved attention hence, not many diagnostic assessments are available for teachers to use in classrooms (Alderson, 2005). More research is needed to construct diagnostic assessment tools to include cognitive tasks suited for measuring learners' strengths and weaknesses on a specific topic. An ongoing formative assessment enables teachers to evaluate their students' learning progress longitudinally. These results then can be an effective tool for teachers to report to the parents as well as school administrators to improve the quality of educational system and to be able to allocate needed resources for the advancement of student's performance at school settings.

An important future direction for CDMs is developing such formative tests for large scaled assessments. An integration of CDM with computerized adaptive testing is a new trend in CDM research. A computerized attribute mastery evaluator can create an immediate feedback for practitioners and students. These feedbacks enable teachers to evaluate their pedagogy, curriculum, and design their approach accordingly.

One possible future direction in computerizing CDM is its potential to customize student's assessments according to their attribute mastery/non-mastery level and to provide customized individual eLearning instruction targeted at those specific sets of attribute by providing an ongoing formative assessment tools with the aim of mastery.

Although there are efforts in combining formative and summative assessments at state level, its implementation has yet to succeed. The Smarter Balanced Assessment Consortium (Smarter Balanced, http://www.smarterbalanced.org/) is one of the existing programs supported by Department of Education aiming at evaluating students' progress. Such programs can provide detailed feedback on students' learning.

The Partnership for Assessment of Readiness for College and Careers (PARCC, http://www.parcconline.org/about-parcc) is another existing consortium working toward development of formative assessment. Similar to Smarter Balanced, they aim at a type of assessment that helps student succeed by providing them with pinpointed and timely feedback to prepare them for college and careers.

Despite recent interest in CDM, there is a pressing need for more research in understanding their behavior under various designs. There is too a call for an evaluation of techniques. This call comes with the hope of identifying and synthesizing from an evaluation of available techniques a more general technique, one more easily implemented and understood by practitioners.

In conclusion, addressing technical questions regarding CDM’s implementation is necessary to ensure their estimation accuracy. It is essential to state that CDMs can be an alternative approach to test analysis with valuable results (Henson, 2009). Therefore, it is important to provide as many studies as possible insofar that collecting and presenting empirical evidence can demonstrate the superiority of CDMs when compared to established testing theories (e.g. IRT and CTT). Future studies of CDMs are important to establish reliability, validity, and generalizability of these models.

For a better understanding of the research process, true, overfited, & opposite Q-Matrix used in simulation design for all dataset were shown in Appendix A and Mplus and SAS coding are included in Appendix B and Appendix C respectively.

References

(Complete References of Both Studies)

Agrestic, A., & Finlay, B. (1997). *Statistical methods for the social science*. Princeton, NJ: Prentice Hall.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics, 30*, 9-14.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* London, England: Continuum.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. Lighton, & M. J. Gierl (Eds.), *Cognitive assessment for education: Theory and applications* (pp. 85-115). New York, NY: Cambridge University.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Bruin, J. (2006). New test: Command to compute new test. *Academic Technology Services, Statistical Consulting Group.* Retrieved from http://www.ats.ucla.edu/stat/stata/ado/analysis/

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 43*(3), 423-466.

Choi, H., Templin, J., & Cohen, A. (2010). *The impact of model misspecification on estimation accuracy in diagnostic classification models.* Unpublished manuscript.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford.

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

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies on cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595-624.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 37*, 359–374.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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

Galeshi, R., & Skaggs, G. E. (2010, May). *Retrofitting TIMSS 2007 mathematics test to cognitive diagnostic model: By Fusion Model.* Paper presented at annual meeting of the Psychometric Society, Athens, GA.

Galeshi, R., & Skaggs, G. (2012). *Item parameter estimation accuracy with a cognitive diagnostic model, CRUM*. Manuscript in preparation.

Gewekei, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica, 57*(6), 1317-1339.

Gierl, M. J., Cui, Y., & Zhou, J. (2008). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(3), 293-313.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (ETS Report RR-08-45). Princeton, NJ: Educational Testing Services.

Haertel, E. H. (1989). Using restricted latent class models to map skill structure of achievement items. *Journal of Educational Measurement, 26,* 301–321.

Harre, R. (1970). *The principles of scientific thinking.* Chicago, IL: University of Chicago.

Harrell, L. M. (2009). *Accuracy of global fit indices as indictors of multidimensionality in multidimensional Rasch analysis* (Doctoral dissertation). Retrieved from http://scholar.lib.vt.edu/theses/etd-search.html

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

Harwell, M. Stone, C. A., Hsu, T. C., & Krisci, L. (1996). *Monte Carlo Methods*, London, England: Methuen.

Henson, R., Templin, J., & Douglas, J. (2007). Use of subscores for estimation of skill masteries. *Journal of Educational Measurement, 44,* 361-376.

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

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Lighton, & M. J. Gierl (Eds.), *Cognitive assessment for education: Theory and applications* (pp. 19-50). New York, NY: Cambridge University.

Hurvich, C.M., & Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76 (2), 297-307.

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

Jiang, H. DiBello, L.V., & Stout, W. (1996). *An estimation procedure for the structural parameters of the unified Cognitive/IRT model*. Retrieved from http://ezproxy.lib.vt.edu:8080/login?url=http://search.proquest.com/docview/62674884?accountid=14826

Junker, B. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*. Talk presented at The Committee on the Foundations of Assessment, National Research Council, University of Pittsburgh, Pittsburgh, PA.

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

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling.* New York, NY: Guilford.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods Research, 33* (2), 188-229. doi: 10.1177/0049124103262065

Kunina-Habenicht, O. (2010). *Theoretical and practical considerations for implementing diagnostic classification models: Insights from simulation-based and applied research*. (Doctoral dissertation, Humboldt University of Berlin). Retrieved from http://edoc.hu-berlin.de/dissertationen

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81.

Lai, H., Gierl, M. J., & Cui, Y. (2012, May). Item consistency index: An index for cognitive diagnostic assessment. *Proceeding of* the *Centre for Research in Applied Measurement and Evaluation*, Vancouver, Canada. Retrieved from http://www2.education.ualberta.ca.ezproxy.lib.vt.edu:8080/educ/psych/crame/docs/April%202012/NCME%202012%20ICI.pdf

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and practice.* (Eds.). New York, NY: Cambridge University.

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

Linn, R. L. (1989). *Has item response theory increased the validity of achievement test scores?* (CSE Report 302). Retrieved from UCLA, Center for Research on Evaluation, Standards, and Student Testing website: http://www.cse.ucla.edu/products/reports/tr302.pdf

Linn, R. L. (1990). Diagnostic testing. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2,* 99-120.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*(4), 523-547.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187-212.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage.

Mislevy, R. J. (1995, May). *Probability-based inference in cognitive diagnosis*. Paper presented at the Office of Naval Research Contractors Conference, Iowa City, IA.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to Evidence-Centered Design*. (Technical Report No. 632). Retrieved from University of California, Center for the Study of Evaluation website: http://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Muthén, L.K., & Muthén, B.O. (2010). Mplus user's guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Norris, S. P., Macnab, J. S., & Phillips, L. M. (2007). Cognitive modeling of performance on diagnostic achievement tests: A philosophical analysis and justification. In J. P. Lighton, & M. J. Gierl (Eds.), *Cognitive assessment for education: Theory and applications*. (275-318). New York, NY: Cambridge University.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit sore scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53* (3), 315-333.

Patz, J. R., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

Qian, S. S., Stow, C. A., & Borsuk, M. E. (2003). On the Monte Carlo methods for Bayesian inference. *Ecological Modeling, 15,* 269-277.

Roussos, L. A., Dibello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion models skills diagnosis system. In J. P. Lighton, & M. J. Gierl (Eds.), *Cognitive assessment for education: Theory and applications*. (275-318). New York, NY: Cambridge University.

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

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

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

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

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52,* 333–343.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Eds.), *Educational measurement* (pp. 263–331). New York, NY: Macmillan.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173-180.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Applied Statistics, 51,* 337-350.

Tatsuoka, K. K., Birenbaum, M., & Arnold, J. (1989). On the stability of students' rules of operation for solving arithmetic problems. *Journal of Educational Measurement, 26*(4), 351-361.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*(4), 901–926.

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

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

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

US. Department of Education (2001). *No Child Left Behind Act of 2001* (Publication No. 107-110). Retrieved from http://www2.ed.gov/nclb/landing.jhtml

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

Weakliem, D. (2004). Introduction to the special issue on model selection. *Sociological Methods and Research 33*, 167-87.

Appendix A: Overfit and Opposite Q-MATRIX

Table A.1

True, Overfited, & Opposite Q-Matrix Used in Simulation Design for Dataset of K=5 & i=31

True Q-Matrix					opposite Q-Matrix					Over fit Q-Matrix				
K1	K2	K3	K4	K5	K1	K2	K3	K4	K5	K1	K2	K3	K4	K5
0	0	0	0	1	1	1	1	1	0	1	1	1	1	1
0	0	0	1	0	1	1	1	0	1	1	1	1	1	1
0	0	0	1	1	1	1	1	0	0	1	1	1	1	1
0	0	1	0	0	1	1	0	1	1	1	1	1	1	1
0	0	1	0	1	1	1	0	1	0	1	1	1	1	1
0	0	1	1	0	1	1	0	0	1	1	1	1	1	1
0	0	1	1	1	1	1	0	0	0	1	1	1	1	1
0	1	0	0	0	1	0	1	1	1	1	1	1	1	1
0	1	0	0	1	1	0	1	1	0	1	1	1	1	1
0	1	0	1	0	1	0	1	0	1	1	1	1	1	1
0	1	0	1	1	1	0	1	0	0	1	1	1	1	1
0	1	1	0	0	1	0	0	1	1	1	1	1	1	1
0	1	1	0	1	1	0	0	1	0	1	1	1	1	1
0	1	1	1	0	1	0	0	0	1	1	1	1	1	1
0	1	1	1	1	1	0	0	0	0	1	1	1	1	1
1	0	0	0	0	0	1	1	1	1	1	1	1	1	1
1	0	0	0	1	0	1	1	1	0	1	1	1	1	1
1	0	0	1	0	0	1	1	0	1	1	1	1	1	1
1	0	0	1	1	0	1	1	0	0	1	1	1	1	1
1	0	1	0	0	0	1	0	1	1	1	1	1	1	1
1	0	1	0	1	0	1	0	1	0	1	1	1	1	1
1	0	1	1	0	0	1	0	0	1	1	1	1	1	1
1	0	1	1	1	0	1	0	0	0	1	1	1	1	1
1	1	0	0	0	0	0	1	1	1	1	1	1	1	1
1	1	0	0	1	0	0	1	1	0	1	1	1	1	1
1	1	0	1	0	0	0	1	0	1	1	1	1	1	1
1	1	0	1	1	0	0	1	0	0	1	1	1	1	1
1	1	1	0	0	0	0	0	1	1	1	1	1	1	1
1	1	1	0	1	0	0	0	1	0	1	1	1	1	1
1	1	1	1	0	0	0	0	0	1	1	1	1	1	1
1	1	1	1	1	0	0	0	0	0	1	1	1	1	1

Table A.2

True, Overfited, & Opposite Q-Matrix Used in Simulation Design for Dataset of K=3 & i=35

	True Q-Matrix			Opposite Q-Matrix			Overfit Q-Matrix		
item	K1	K2	K3	K1	K2	K3	K1	K2	K3
1	0	0	1	1	1	0	1	1	1
2	0	1	0	1	0	1	1	1	1
3	0	1	1	1	0	0	1	1	1
4	1	0	0	0	1	1	1	1	1
5	1	0	1	0	1	0	1	1	1
6	1	1	0	0	0	1	1	1	1
7	1	1	1	0	0	0	1	1	1
8	0	0	1	1	1	0	1	1	1
9	0	1	0	1	0	1	1	1	1
10	0	1	1	1	0	0	1	1	1
11	1	0	0	0	1	1	1	1	1
12	1	0	1	0	1	0	1	1	1
13	1	1	0	0	0	1	1	1	1
14	1	1	1	0	0	0	1	1	1
15	0	0	1	1	1	0	1	1	1
16	0	1	0	1	0	1	1	1	1
17	0	1	1	1	0	0	1	1	1
18	1	0	0	0	1	1	1	1	1
19	1	0	1	0	1	0	1	1	1
20	1	1	0	0	0	1	1	1	1
21	1	1	1	0	0	0	1	1	1
22	0	0	1	1	1	0	1	1	1
23	0	1	0	1	0	1	1	1	1
24	0	1	1	1	0	0	1	1	1
25	1	0	0	0	1	1	1	1	1
26	1	0	1	0	1	0	1	1	1
27	1	1	0	0	0	1	1	1	1
28	1	1	1	0	0	0	1	1	1
29	0	0	1	1	1	0	1	1	1
30	0	1	0	1	0	1	1	1	1
31	0	1	1	1	0	0	1	1	1
32	1	0	0	0	1	1	1	1	1
33	1	0	1	0	1	0	1	1	1
34	1	1	0	0	0	1	1	1	1
35	1	1	1	0	0	0	1	1	1

Table A.3

True, Opposite, &, Overfited Q-Matrix Used in Simulation Design for Dataset of K=3 & i=7

	True Q-Matrix			Opposite Q-Matrix			Overfit Q-Matrix		
item	K1	K2	K3	K1	K2	K3	K1	K2	K3
1	0	0	1	1	1	0	1	1	1
2	0	1	0	1	0	1	1	1	1
3	0	1	1	1	0	0	1	1	1
4	1	0	0	0	1	1	1	1	1
5	1	0	1	0	1	0	1	1	1
6	1	1	0	0	0	1	1	1	1
7	1	1	1	0	0	0	1	1	1

Appendix B: Sample of SAS Codes for Data Generation

```
%macro simulation;
%let num_students=50;
%let num_skills=5;
/* Activate to send log to external file if needed;*/
PROC PRINTTO log = "C:\cdm1\LOG\&log..txt" NEW; run;

data nu1;
        array nu_num{&num_total} nu_num1-nu_num&num_total;
        array nu{&num_total} nu1-nu&num_total;
        %do i=1 %to &num_total;
                nu_num{&i}=exp(&&mu&i);
        %end;
        nu_denom=sum(of nu_num1-nu_num&num_total);
        %do j=1 %to &num_total;
                nu{&j}=nu_num{&j}/nu_denom;
        %end;
        drop nu_denom nu_num1-nu_num&num_total;
run;

proc transpose data=nu1 out=nu2; run;

data nu;
        set nu2(rename=(col1=nu));
        combination=_n_;
        drop _name_;
run;

data all_skills;
        array skills{&num_skills} skills1-skills&num_skills;
        do j=1 to 2;
        do i=1 to &num_skills;
                if j=1 then skills{i}=1;
                else skills{i}=0;
                if skills{j}=. then skills{j}=0;
        end;
        output;
        end;
        drop i j;
run;

proc means data=all_skills completetypes;
        class skills1-skills&num_skills;
        output out=all_skills1;
run;
```

```
/* ADDED: To change the format of Class_membership to ascii file*/
data _NULL_;
set Class_membership;
file "C:\CDM1\Class_membership..dat";
put  nu class;
run;

/*continue coding*/
proc sort data=simulated_data; by student_num combinations;run;
proc transpose data=simulated_data(keep=student_num combinations question x_ic)
          out=student&seed(drop=_name_) prefix=question;
          id question;
          by student_num combinations;
run;

* CHANGED Import original data into work library, save as .dat file for Mplus;
data data&num_reps;
          set student&seed;
          file "C:\CDM1\&datafile..dat";
          put  student_num Combinations question1-question&numitem ;
RUN;

DATA Q-Matrix;
          set Q_Matrix;
/* Renaming attributes to common format*/
          array skills{&num_skills} skills1-skills&num_skills;
          ARRAY new(&num_skills) itematt1-itematt&num_skills;
                    DO i=1 TO &num_skills; new(i)=skills(i); END;
item= _N_;
DROP i skills1-skills&num_skills;
run;

DATA _NULL_;
SET Q-MATRIX;
FILE "C:\CDM1\Q-MATRIX.dat";
put itematt1 itematt2 itematt3 itematt4 itematt5;
run;

%end;
%mend;
%simulation;
```

Appendix C: Sample of SAS Codes for MPLUS Input

```
/*Create a command File for Mplus*/
data _null_;
file "c:\cdm\&commandfile..inp";
put "TITLE: CRUM with only main Effect";
put "DATA: FILE IS c:\CDM\&datafile..dat;"; **** modified ****;
put "VARIABLE:  NAMES= STUDENT CLASS X1-X7;";
PUT "USEVARIABLE = x1-x7;";
PUT "CATEGORICAL = x1-x7;";
PUT "CLASSES = c(8);";

PUT "ANALYSIS:";
PUT "TYPE=MIXTURE; !estimates latent classes;";
PUT "STARTS=0; !turn off multiple random start feature (disabled anyway);";

put "MODEL:";
PUT '%OVERALL%';
PUT "[C#1] (m1); !latent variable mean for attribute pattern [0,0,0];";
PUT "[C#2] (m2); !latent variable mean for attribute pattern [0,0,1];";
PUT "[C#3] (m3); !latent variable mean for attribute pattern [0,1,0];";
PUT "[C#4] (m4); !latent variable mean for attribute pattern [0,1,1];";
PUT "[C#5] (m5); !latent variable mean for attribute pattern [1,0,0];";
PUT "[C#6] (m6); !latent variable mean for attribute pattern [1,0,1];";
PUT "[C#7] (m7); !latent variable mean for attribute pattern [1,1,0];";

put '%c#1% !for attribute pattern [0,0,0];';
put "[x1$1] (t1_1); !threshold for item 1 LCDM kernel 1";
put "[x2$1] (t2_1); !threshold for item 2 LCDM kernel 1";
put "[x3$1] (t3_1); !threshold for item 3 LCDM kernel 1";
put "[x4$1] (t4_1); !threshold for item 4 LCDM kernel 1";
put "[x5$1] (t5_1); !threshold for item 5 LCDM kernel 1";
put "[x6$1] (t6_1); !threshold for item 6 LCDM kernel 1";
put "[x7$1] (t7_1); !threshold for item 7 LCDM kernel 1";

put '%c#2% !for attribute pattern [0,0,1];';
put "[x1$1] (t1_2); !threshold for item 1 LCDM kernel 1";
put "[x2$1] (t2_1); !threshold for item 2 LCDM kernel 1";
put "[x3$1] (t3_2); !threshold for item 3 LCDM kernel 2";
put "[x4$1] (t4_1); !threshold for item 4 LCDM kernel 1";
put "[x5$1] (t5_2); !threshold for item 5 LCDM kernel 2";
put "[x6$1] (t6_1); !threshold for item 6 LCDM kernel 2";
put "[x7$1] (t7_2); !threshold for item 7 LCDM kernel 2";

put '%c#3% !for attribute pattern [0,1,0];';
put "[x1$1] (t1_1); !threshold for item 1 LCDM kernel 1";
put "[x2$1] (t2_2); !threshold for item 2 LCDM kernel 2";
put "[x3$1] (t3_3); !threshold for item 3 LCDM kernel 1";
put "[x4$1] (t4_1); !threshold for item 4 LCDM kernel 2";
put "[x5$1] (t5_1); !threshold for item 5 LCDM kernel 1";
put "[x6$1] (t6_2); !threshold for item 6 LCDM kernel 3";
put "[x7$1] (t7_3); !threshold for item 7 LCDM kernel 3";

put '%c#4% !for attribute pattern [0,1,1];';
put "[x1$1] (t1_2); !threshold for item 1 LCDM kernel 1";
put "[x2$1] (t2_2); !threshold for item 2 LCDM kernel 2";
put "[x3$1] (t3_4); !threshold for item 3 LCDM kernel 2";
put "[x4$1] (t4_1); !threshold for item 4 LCDM kernel 2";
put "[x5$1] (t5_2); !threshold for item 5 LCDM kernel 2";
put "[x6$1] (t6_2); !threshold for item 6 LCDM kernel 4";
put "[x7$1] (t7_4); !threshold for item 7 LCDM kernel 4";

put '%c#5% !for attribute pattern [1,0,0];';
put "[x1$1] (t1_1); !threshold for item 1 LCDM kernel 2";
put "[x2$1] (t2_1); !threshold for item 2 LCDM kernel 1";
put "[x3$1] (t3_1); !threshold for item 3 LCDM kernel 1";
put "[x4$1] (t4_2); !threshold for item 4 LCDM kernel 3";
put "[x5$1] (t5_3); !threshold for item 5 LCDM kernel 3";
put "[x6$1] (t6_3); !threshold for item 6 LCDM kernel 1";
put "[x7$1] (t7_5); !threshold for item 7 LCDM kernel 5";

put '%c#6% !for attribute pattern [1,0,1];';
put "[x1$1] (t1_2); !threshold for item 1 LCDM kernel 2";
put "[x2$1] (t2_1); !threshold for item 2 LCDM kernel 1";
put "[x3$1] (t3_2); !threshold for item 3 LCDM kernel 2";
put "[x4$1] (t4_1); !threshold for item 4 LCDM kernel 3";
put "[x5$1] (t5_4); !threshold for item 5 LCDM kernel 4";
put "[x6$1] (t6_3); !threshold for item 6 LCDM kernel 2";
put "[x7$1] (t7_6); !threshold for item 7 LCDM kernel 6";


put '%c#7% !for attribute pattern [1,1,0];';
put "[x1$1] (t1_1); !threshold for item 1 LCDM kernel 2";
put "[x2$1] (t2_2); !threshold for item 2 LCDM kernel 2";
put "[x3$1] (t3_3); !threshold for item 3 LCDM kernel 1";
put "[x4$1] (t4_2); !threshold for item 4 LCDM kernel 4";
put "[x5$1] (t5_3); !threshold for item 5 LCDM kernel 3";
put "[x6$1] (t6_4); !threshold for item 6 LCDM kernel 3";
put "[x7$1] (t7_7); !threshold for item 7 LCDM kernel 7";

put '%c#8% !for attribute pattern [1,1,1];';
put "[x1$1] (t1_2); !threshold for item 1 LCDM kernel 2";
put "[x2$1] (t2_2); !threshold for item 2 LCDM kernel 2";
put "[x3$1] (t3_4); !threshold for item 3 LCDM kernel 2";
put "[x4$1] (t4_2); !threshold for item 4 LCDM kernel 4";
put "[x5$1] (t5_4); !threshold for item 5 LCDM kernel 4";
put "[x6$1] (t6_4); !threshold for item 6 LCDM kernel 4";
put "[x7$1] (t7_8); !threshold for item 7 LCDM kernel 8";

Put "MODEL CONSTRAINT: !used to define LCDM parameters and constraints";
put "!NOTE: Mplus uses P(X=0) rather than P(X=1) so terms must be multiplied by -1";
put "!One attibute measured: 1 intercept; 1 main effect";

put"!ITEM 1:"
Put "!Q-matrix Entry [0 0 1]";
put "NEW(l1_0 l1_11);    !define LCDM parameters present for item 1";
put "t1_1=-(l1_0);       !set equal to intercept by LCDM kernel";
put "t1_2=-(l1_0+l1_11); !set equal to intercept plus main effect for attribute 1";
put "l1_11>0;            !make sure main effect is positive (higher probability for mastering";

put "!ITEM 2:";
Put "!Q-matrix Entry [0 1 0]";
put "!One attibute measured: 1 intercept; 1 main effect";
put "NEW(l2_0 l2_12);    !define LCDM parameters present for item 2";
put "t2_1=-(l2_0);";
put "t2_2=-(l2_0+l2_12);";
put "l2_12>0;        !the order constraints necessary for the main effect";

put "!ITEM 3:";
put "!Q-matrix Entry [0 1 1]";
put "!two attibutes measured: 1 intercept; 2 main effects;";
put "NEW(l3_0 l3_12 l3_13);       !define LCDM parameters present for item 6";
put "t3_1=-(l3_0);";
put "t3_2=-(l3_0+l3_12);";
put "t3_3=-(l3_0+l3_13);";
put "t3_4=-(l3_0+l3_12+l3_13);";
put "l3_12>0;";
put "l3_13>0;";

put " !ITEM 4:";
put "!Q-matrix Entry [1 0 0]";
put "!two attibutes measured: 1 intercept; 2 main effects; 1 two-way interaction";
put "NEW(l4_0 l4_11);       !define LCDM parameters present for item 4";
put "t4_1=-(l4_0);";
put "t4_2=-(l4_0+l4_11);";
put "l4_11>0;              !the order constraints necessary for the main effects";
put "!ITEM 5:";
put "!Q-matrix Entry [1 0 1]";
put "!two attibutes measured: 1 intercept; 2 main effects; 1 two-way interaction";
put "NEW(l5_0 l5_11 l5_13);       !define LCDM parameters present for item 5";
put "t5_1=-(l5_0);";
put "t5_2=-(l5_0+l5_11); ";
put "t5_3=-(l5_0+l5_13);";
put "t5_4=-(l5_0+l5_11+l5_13); ";
put "l5_11>0;                        !the order constraints necessary for the main effects";
put "l5_13>0;";

put "!ITEM 6:";
put "!Q-matrix Entry [1 1 0]";
put "!two attibutes measured: 1 intercept; 2 main effects; 1 two-way interaction ";
put "NEW(l6_0 l6_11 l6_12);       !define LCDM parameters present for item 6";
put "t6_1=-(l6_0);";
put "t6_2=-(l6_0+l6_11);";
put "t6_3=-(l6_0+l6_11);";
put "t6_4=-(l6_0+l6_11+l6_12);";
put "l6_11>0;";
put "l6_12>0;";


put "!ITEM 7:";
put "!Q-matrix Entry [1 1 1]";
put "!two attibutes measured: 1 intercept; 3 main effects; 3 two-way interactions; 1 three-way";
put "NEW(l7_0 l7_11 l7_12 l7_13);   !define LCDM parameters presen";
put "t7_1=-(l7_0);";
put "t7_2=-(l7_0+l7_13);";
put "t7_3=-(l7_0+l7_12);";
put "t7_4=-(l7_0+l7_12+l7_13);";
put "t7_5=-(l7_0+l7_11);";
put "t7_6=-(l7_0+l7_11+l7_13);";
put "t7_7=-(l7_0+l7_11+l7_12);";
put "t7_8=-(l7_0+l7_11+l7_12+l7_13);";
put "l7_11>0;                      !the order constraints necessa";
put "l7_12>0;";
put "l7_13>0;";

put "OUTPUT:";
put "TECH10; !request additional model fit statistics be reported";

put "SAVEDATA:";
put "FORMAT IS f10.5;            !format for output file";
put "FILE IS c:\cdm\&outputfile..dat; !print attribute estimates for respondents in file list";
put "SAVE = CPROBABILITIES;       !instruct Mplus to save posterior probabilities of class";
```